DOCUMENT RESUME

ED 177 958 HE 011 813

AUTHOR	Jones, Deborah L., Comp.
TITLE	Proceedings: Eighteenth Annual Conference on Research in Medical Education, Washington, D.C., November 6-7, 1979.
INSTITUTION	Association of American Medical Colleges, Washington, D. C.
PUB DATE	Nov 79
NOTE	454p.
EDRS PRICE	MF01/PC19 Plus Postage.
DESCRIPTORS	Certification; Clinical Experience; Clinic Personnel (School); Effective Teaching; Faculty Development; Faculty Evaluation; Grading; *Graduate Medical Education; *Graduate Medical Students; Medical Care Evaluation; *Medical Education; *Medical Evaluation; *Medical Research; Medical Students; Participant Satisfaction; Patients (Persons); Professional Development; Rating Scales; Simulation; Stress Variables; Student Development; Testing
IDENTIFIERS	*Medical School Faculty

ABSTRACT

The Research in Medical Education (RIME) Program Planning Committee's selections for program materials for the eighteenth annual conference on Research in Medical Education are contained in this volume. The agenda consisted of poster sessions, presentation of papers, and presentation of symposia. Poster sessions examined such topics as student development, medical faculty, instruction, and curriculum. Recertification, residency training, improved medical education, and clinical performance were some of the areas discussed in the symposia presentations. Papers presented were: "Patient Simulators in Clinical Teaching" (Donald A. Bosshart); "Assessment of Clinical Reasoning" (Carl F. Hinz, Jr.); "Professional Development" (H. Tristram Engelhardt, Jr.); "Performance Evaluation in Graduate Medical Education" (Fredric D. Burg); "Prediction and Selection" (Paul R. Elliott); "Evaluation of Teaching" (Arthur I. Rothman); "Clinical Performance Rating Scales" (Charles P. Friedman); "Meaning, Communication, and Memory" (Howard L. Stone); "Enhancing Clinical Teaching" (John Casbergue); "Assessment in Continuing Medical Education" (Oscar A. Thorup); "Student Satisfaction and Stress" (C. Warner Johnson); and "Issues in Testing and Grading" (Jayne Middleton). (LC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED177958

PROCEEDINGS

EIGHTEENTH ANNUAL CONFERENCE
ON
RESEARCH IN MEDICAL EDUCATION

U S DEPARTMENT OF HEALTH.
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

AAMC

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

HE 011 893

Sponsored by
ASSOCIATION OF AMERICAN MEDICAL COLLEGES
Division of Educational Measurement and Research
in conjunction with the 90th Annual Meeting
November, 1979

HE011 813

EIGHTEENTH ANNUAL CONFERENCE

ON

RESEARCH IN MEDICAL EDUCATION

Washington, D.C.
Washington Hilton Hotel
November 6-7, 1979

The Eighteenth Annual Conference on Research in Medical Education (RIME) is planned and organized by the RIME Program Planning Committee, which is a committee of the AAMC Group on Medical Education (GME). The material contained in this volume represents the selections of the Planning Committee from submissions made in conformance with the annual Call for Programs material.

PLANNING COMMITTEE
FOR THE
EIGHTEENTH ANNUAL CONFERENCE ON RESEARCH IN MEDICAL EDUCATION

CHAIRMAN: T. Joseph Sheehan, Ph.D.	University of Connecticut
Gary M. Arsham, M.D., Ph.D.	Pacific Medical Center
Richard M. Caplan, M.D.	University of Iowa
Kaaren I. Hoffman, Ph.D.	University of Southern California
Victor R. Neufeld, M.D.	McMaster University
Frank T. Stritter, Ph.D.	University of North Carolina
James B. Erdmann, Ph.D., Secretary	Association of American Medical Colleges

Proceedings Compiled By: Deborah L. Jones

Eighteenth Annual

CONFERENCE ON RESEARCH IN MEDICAL EDUCATION

November 6-7, 1979
Washington Hilton Hotel
Washington, D.C.

POSTER SESSIONS

November 6, 1979
Exhibit Hall
1:00 - 4:00 p.m.

Three Years of Faculty Development Perspectives of the Past, Present, and Future
Sylvia Shellenberger, Ph.D., T. Walter Treadwell, M.D., N. Kay Mashburn, M.Ed.

Developing a Student Data Base for Institutional Research and Evaluation
Claudia M. Cheatham, Carol Q. Porter, Charles P. Friedman

A Method for Predicting National Board: Part I Scores
William E. Sorlie, Diane Essex, Ph.D.

Assessment of the Ability to Observe Affective Behavior in Medical Interviews
Robert B. Cowan, Jr., M.D., Ronald A. Krol, Ed.D., George M. Curtis, M.D.

Use of Multiple Perspectives in Determining Curriculum or Test Content
Reed G. Williams, John H. Shatzer, Janet McDole

Evaluation Strategies for Residency Training in Behavioral Pediatrics, or, "Overcoming Organizational Hurdles"
Pamela J. Trent, Ph.D., Robert A. Hock, Ph.D., W. Sam Yancy, M.D.

Using Confidence Testing in a Physical Diagnosis Course
Robert M. Rippey, Ph.D., Anthony E. Voytovich, M.D.

PRESENTATION OF PAPERS

PATIENT SIMULATORS IN CLINICAL TEACHING

MODERATOR: Donald A. Bosshart, Ed.D.
Northeastern Ohio Universities

Evaluation of a Self-Instructional Training Program For Simulated Patients
Barbara F. Johnson, Ph.D., Linda Perkowski, M.S., Roger A. Lanier, Ph.D.

An Instructional Program Using Patient Instructors as Teachers and Evaluators
Paula Stillman, M.D., Jane Ruggill, Paul Rutala, M.D., Darrell Sabers, Ph.D.

A Comparison of Instructional Techniques: Internal-External Vs. External Only
Walter Gerber, M.D., Sandra Matthes, M.A.T., Mark Albanese, M.A.

DATE: November 7, 1979
TIME: 8:30 - 10:15 a.m LOCATION: Military

ASSESSMENT OF CLINICAL REASONING

MODERATOR: Carl F. Hinz, Jr., M.D.
University of Connecticut

The Underlying Structure of Clinical Problem-Solving: Process or Content?
P.H. Harasym, F. Alexander, J.S. Baumber, H. Bryant, D. Fundytus, I. MacPhail, R. Preshaw, M. Sosnowski, M. Watanabe, G. Wyse

The "Triple Jump" Exercise - A Structured Measure of Problem Solving and Self Directed Learning
C. Painvin, V. Neufeld, G. Norman, I. Walker, G. Whelan

Evaluating Analytic Reasoning in the Management of the Multiproblem Patient: A System for Attaining Reliability and Validity
A. E. Voytovich, M.D., L. Harper, M.D., R. M. Rippey, Ph.D.

DATE: November 7, 1979
TIME: 8:30 - 10:15 a.m. LOCATION: Hemisphere

PRESENTATION OF PAPERS

PROFESSIONAL DEVELOPMENT

MODERATOR: H. Tristram Engelhardt, Jr., M.D., Ph.D.
Georgetown University

Assessing the Moral Development of Medical Students: An Empirical Study
M. Harry Daniels, Ph.D., George L. Baker, M.D.

Moral Reasoning and Physicians' Decisions in Cases of Critical Illness
Daniel Candee, Ph.D., T. Joseph Sheehan, Ph.D., Charles D. Cook, M.D., Susan R. Husted, B.S.

Coping and Ego Development in Internship
Julie C. Donnelly, Ed.D.

DATE: November 7, 1979
TIME: 8:30 - 10:15 a.m. LOCATION: Thoroughbred

PERFORMANCE EVALUATION IN GRADUATE MEDICAL EDUCATION

MODERATOR: Fredric D. Burg, M.D.
National Board of Medical Examiners

Utah Continuing Studies on Physician and Physician-in-Training Performance
Dominic Albo, M.D., Calvin W. Taylor, Ph.D., Philip Price, M.D., Brent Page, M.S.

Failure of Examination Answers to Evaluate Actual Practice Patterns by Medical Housestaff in an Outpatient Clinic
Steven D. Rose, M.D., Lourdes C. Corman, M.D., John Robbins, M.D.

Reliability in Follow-up Evaluations of Graduates as Residents
Rosalia E. A. Paiva, Ph.D.

The Effects of Clinically Relevant Multiple-Choice Items on the Statistical Discrimination of Physician Clinical Competence
Steven M. Downing, Ph.D., Jack L. Maatsch, Ph.D.

DATE: November 7, 1979
TIME: 10:30 a.m. - 12:30 p.m. LOCATION: Military

PRESENTATION OF PAPERS

PREDICTION AND SELECTION

MODERATOR: Paul R. Elliott, Ph.D.
Florida State University

The Relationship Between Scores on the New Medical College Admission Test and Performance in the First Year of Medical School: A Preliminary Study
Bruce C. Deighton, Douglas U. Smith, Ph.D., Richard E. Gallagher, Ph.D.

A Factor Analytic Study of the Old and New MCAT Examinations
John B. Molidor, Ph.D., Arthur S. Elstein, Ph.D.

A Non-Linear Discriminant Analytic Approach to Predicting Medical Student Performance
Charles P. Friedman, Claudia M. Cheatham, Carol Q. Porter, William E. Bakewell

A Structured Interview for the Selection of Physician's Assistant Students
Bruce R. Niebuhr, Ph.D., Steven R. Shelton, P.A.-C., Shirley A. Werner, P.A.-C.

DATE: November 7, 1979
TIME: 10:30 a.m. - 12:30 p.m. LOCATION: Hemisphere

EVALUATION OF TEACHING

MODERATOR: Arthur I. Rothman, Ed.D.
University of Toronto

How Medical Students Use Objectives
Terrill A. Mast, Ph.D., George P. Evans, M.S., Reed G. Williams, Ph.D., David L. Silber, M.D.

What Should Students Evaluate? A Comparison of Faculty and Student Perceptions
Michael B. Donnelly, Ph.D., Carolyn Puhl, Daniel S. Fleisher, M.D., Jeffrey H. Garrison, M.D., Stanley Kaplan, Ph.D.

Teaching Effectiveness: A Comparison of Student Ratings and Graduate Effectiveness
Roger A. Lanier, Ph.D., Byron Williams, P.A.-C., Bruce R. Niebuhr, Ph.D.

Comparison of Faculty and Chairmen Perceptions About Promotion Criteria
Craig Gjerde, Ph.D., Sandra Colombo, M.A.

DATE: November 7, 1979
TIME: 10:30 a.m. - 12:30 p.m. LOCATION: Thoroughbred

PRESENTATION OF PAPERS

CLINICAL PERFORMANCE RATING SCALES

MODERATOR: Charles P. Friedman, Ph.D.
University of North Carolina

The Development of Rating Scales to Assess the Clinical Performance of Medical Students
Vivian Erviti, Ph.D., Lawrence J. Fabrey, M.S., Jane V. Bunce

Inter-Rater Reliability of Clinical Performance Ratings
T. E. Dielman, Ph.D., Alan L. Hull, M.S., Wayne K. Davis, Ph.D.

The Relationship Between Reported Activities and Clinical Evaluations of Third-Year Medical Students in a Medicine Clerkship
Alan L. Hull, M.S., Judith G. Calhoun, Ph.D., Wayne K. Davis, Ph.D.

DATE: November 7, 1979
TIME: 1:30 - 3:30 p.m. LOCATION: Military

MEANING, COMMUNICATION, AND MEMORY

MODERATOR: Howard L. Stone, Ph.D.
University of Wisconsin

The Communication of Uncertainty
G. D. Bryant, G. R. Norman

Physician-Patient Communication: Interpretations of Non-Technical Phrases
Marcia Z. Wile, Ph.D., Juanita R. Wingenfeld, B.A., Charlene Ratner, M.S., Jane E. Mesnard, B.S., Lois P. Mack, M.S., Brian A. Kastner, D.D.S., James R. Hickey, M.S., Steven Dickoff, B.S.

Clinical Experience and the Structure of Memory
G. R. Norman, L. L. Jacoby, J. W. Feightner, E. J. M. Campbell

DATE: November 7, 1979
TIME: 1:30 - 3:30 p.m. LOCATION: Hemisphere

PRESENTATION OF PAPERS

ENHANCING CLINICAL TEACHING

MODERATOR: John Casbergue, Ph.D.
Michigan State University

Influence of Sequencing in Clinical Tutorials for Second Year Medical Students
C.C. Clawson, M.D., Ilene Harris, Ph.D., Judith Garrard, Ph.D.

Development, Implementation and Evaluation of a Program to Improve Clinical Teaching
Mark I. Patridge, Ilene B. Harris, Robert A. Petzel

Verbal Participation and Outcomes in Medical Education: A Study of Third-year Clinical Discussion Groups
Patricia Joan Foster, Ph.D.

Written vs Oral Feedback: Their Effect on Learning in an Internal Medicine Clerkship
John F. Markus, M.S., Terrill A. Mast, Ph.D., Norman G. Soler, M.D.

DATE: November 7, 1979
TIME: 1:30 - 3:30 p.m. LOCATION: Thoroughbred

ASSESSMENT IN CONTINUING MEDICAL EDUCATION

MODERATOR: Oscar A. Thorup, M.D.
University of Virginia

Using Audit as a Needs Assessment and Evaluation for a Continuing Education Program
Sharon E. Paino, R.N., David W. Cline, M.D., Lawrence Demarest, Ph.D.

Continuing Medical Education: An Assessment of Impact Upon Neonatal Care in Community Hospitals
David D. Wirtschafter, M.D., James Sumners, M.D., Ed Colvin, M.D., Eric Knox, M.D., Malcolm Turner, Ph.D., George Cassady, M.D., Pat Beale, R.N., Margaret Budrich, R.N., C. Michael Brooks, Ed.D.

Medical Care Evaluation: An Experience in Continuing Medical Education
L.J. Sandlow, M.D., P.G. Bashook, Ed.D., J. Maxwell, M.A.

Development and Evaluation of a Community-Based Pulmonary Education System
Roland G. Hiss, M.D., Jeoffrey K. Stross, M.D., Wayne K. Davis, Ph.D., Roderick Macdonald, M.A.

Physician Practice Profiles
R. Wayne Putnam, M.D., Lynn Curry, Ph.D.

DATE: November 7, 1979
TIME: 3:45 - 5:45 p.m. LOCATION: Military

PRESENTATION OF PAPERS

STUDENT SATISFACTION AND STRESS

MODERATOR: C. Warner Johnson, M.D.
University of Southern California

Identifying Satisfying and Stressful Events in Medical School
Richard Maisiak, Ph.D., James Britton, III, and Ronald Meredith, Psy.D.

Medical Vs. Graduate Students: Socialization and Stress
Joanna Spiro, Ed.D., Marcella Roenneburg, Betty Joan Haly, M.D.

Dogmatism and Values Orientation of Residents in Family Practice
Kenneth J. Wade, Ph.D., Roland P. Ficken, Ph.D.

DATE: November 7, 1979
TIME: 3:45 - 5:45 p.m. LOCATION: Hemisphere

ISSUES IN TESTING AND GRADING

MODERATOR: Jayne Middleton, Ed.D.
University of Kentucky

The Effects of Commercial Tutoring on Medical Licensure Examinations
Lowell K. Scott, Ed. D., C.W. Scott, M.D., Russell D. Cunningham, M.D., Paul A. Palmisano, M.D., Nass J. Cannon, M.D., Sam Brown, Ed.D.

Curriculum Bias: A Study of a College-Certifying System for Students at Two Basic Medical Science Schools
Diane L. Essex, William E. Sorlie

Medical Students Learning Processes and Their Relation to Test Performance and Attitudes Toward Item Formats
Christine Nu Viet Vu, Ph.D.

DATE: November 7, 1979
TIME: 3:45 - 5:45 p.m. LOCATION: Thoroughbred

PRESENTATION OF SYMPOSIA

TOPIC: New Experiments to Improve Premedical and Medical Education

Chairman: John I. Sandson, M.D.

Participants: Ernest Blaustein, Ph.D.
Peter A. Stewart, Ph.D.
Arnold W. Ravin, Ph.D.
Miriam B. Rock, Ph.D.

Discussant: Mark F. Jackman, Ph.D.

DATE: November 7, 1979
TIME: 8:30 - 10:15 a.m. LOCATION: Cabinet

TOPIC: Research in Residency Training For Primary Care

Chairman: Edwin F. Rosinski, Ed.D.

Participants: Edwin F. Rosinski, Ed.D.
Janet T. Pozen, Ph.D.
Evelyn R. Dienst, Ph.D.
Leighton E. Cluff, M.D.

DATE: November 7, 1979
TIME: 8:30 - 10:15 a.m. LOCATION: Conservatory

PRESENTATION OF SYMPOSIA

TOPIC: United States Citizen Foreign-Trained Medical Students Accepted With Advanced Standing Into American Medical Schools: Background, Reentry and Prognosis

Organizer: Paula L. Stillman, M.D.

Chairman: Ethel Weinberg, M.D.

Participants: Lester M. Geller, Ph.D.
Paula L. Stillman, M.D.
Jane S. Ruggill, M.S.W.
Paul J. Rutala, M.D.
Sarah M. Dinham, Ph.D.
Darrell L. Sabers, Ph.D.
Barry Stimmel, M.D.

DATE: November 7, 1979
TIME: 8:30 - 10:15 a.m.
LOCATION: Jefferson East

TOPIC: Results of a National Study of a Computer Simulation (CASE) and The American Board of Internal Medicine's Recertification Examination: THE MERIT PROJECT

Chairman: George D. Webster, M.D.

Participants: Lynn O. Langdon, M.A.
John A. Meskauskas, M.S.
John J. Norcini, B.A.
George D. Webster, M.D.

Discussants: Arthur S. Elstein, Ph.D.
Victor K. Neufeld, M.D.

DATE: November 7, 1979
TIME: 10:30 a.m. - 12:30 p.m.
LOCATION: Cabinet

PRESENTATION OF SYMPOSIA

TOPIC: Perspectives on Humanistic Medical Education

Chairman: Richard Fletcher, Ph.D.

Participants: Robert T. Manning, M.D.
Cindy Taylor
Gerald Holman, M.D.
Ronald Jordan, Ph.D.

DATE: November 7, 1979
TIME: 10:30 a.m. - 12:30 p.m. LOCATION: Conservatory

TOPIC: Research on Women Physicians: Recent Empirical Findings and Areas for Future Investigation

Chairman: Dorothy Rosenthal Mandelbaum, Ph.D.

Participants: Lillian Kaufman Cartwright, Ph.D.
Sandra L. Chaff
Marilyn Heins, M.D.
Dorothy Rosenthal Mandelbaum, Ph.D.

DATE: November 7, 1979
TIME: 10:30 a.m. - 12:30 p.m. LOCATION: Jefferson East

PRESENTATION OF SYMPOSIA

TOPIC: Medical Problem Solving: A Reanalysis and Alternative Perspectives

Chairman: William C. McGaghie, Ph.D.

Participants: William C. McGaghie, Ph.D.
Anthony LaDuca, Ph.D.
Eta Berner, Ed.D.
John D. Engel, Ph.D.

Discussants: Arthur S. Elstein, Ph.D.
Sarah A. Sprafka, Ph.D.

DATE: November 7, 1979
TIME: 1:30 - 3:30 p.m. LOCATION: Cabinet

TOPIC: Alternatives to Traditional Methods of Curriculum Evaluation

Chairman: Paul L. Grover, Jr., Ph.D.

Participants: Paul L. Grover, Jr., Ph.D.
Judith Sackoff, M.S.
Gerry R. Schermerhorn, M.A.
Michael M. Ravitch, Ph.D.
Ronald W. Richards, Ph.D.

Discussant: Henry B. Slotnick, Ph.D.

DATE: November 7, 1979
TIME: 1:30 - 3:30 p.m. LOCATION: Conservatory

PRESENTATION OF SYMPOSIA

TOPIC: New Horizons in Cost Containment Curricula

Chairman: Mohan L. Garg, Sc.D.

Participants: Michael J. Garland, Ph.D.
William B. Applegate, M.D.
John M. Eisenberg, M.D.

Discussants: John G. Freymann, M.D.
James I. Hudson, M.D.

DATE: November 7, 1979
TIME: 1:30 - 3:30 p.m. LOCATION: Jefferson East

TOPIC: Patient/Health Education: Training For What?

Chairman: Gary M. Arsham, M.D., Ph.D.

Participants: Gary M. Arsham, M.D., Ph.D.
Edward E. Bartlett, M.P.H.
Ellen J. Cohen, Ph.D.
Wendy Diane Squyres, Ph.D

Discussant: Merlin K. DuVal, M.D.

DATE: November 7, 1979
TIME: 3:45 - 5:45 p.m. LOCATION: Cabinet

PRESENTATION OF SYMPOSIA

TOPIC: The Clinical Performance of Medical School Graduates: Theoretical, Methodological, and Empirical Perspectives

Organizer: Louise Arnold, Ph.D.

Chairman: Jerry A. Royer, M.D.

Participants:
Sarah M. Dinham, Ph.D.
Christel A. Woodward, Ph.D.
Louise Arnold, Ph.D.
Jonathan Keck, Ph.D.
T. Lee Willoughby
Virginia Calkins
Jerry A. Royer, M.D.
John C. Reid, Ph.D.
Jon Veloski

Discussant: Rosalia E. A. Paiva, Ph.D.

DATE: November 7, 1979
TIME: 3:45 - 5:45 p.m.
LOCATION: Conservatory

TOPIC: The Role of Decision Analysis in Medical Education and Practice

Chairman: G. Anthony Gorry, Ph.D.

Participants:
Jerome P. Kassirer, M.D.
Harvey V. Fineberg, M.D.
Stephen G. Pauker, M.D.
Arthur S. Elstein, Ph.D.
John C. Peirce, M.D.

DATE: November 7, 1979
TIME: 3:45 - 5:45 p.m.
LOCATION: Jefferson East

TABLE OF CONTENTS

POSTER SESSIONS

PRECIS. 1

Three Years of Faculty Development Perspectives of the Past, Present, and Future. 3

Developing a Student Data Base for Institutional Research and Evaluation. 9

A Method for Predicting National Board: Part I Scores. 14

Assessment of the Ability to Observe Affective Behavior in Medical Interviews. 20

Use of Multiple Perspectives in Determining Curriculum or Test Content. 26

Evaluation Strategies for Residency Training in Behavioral Pediatrics, or, "Overcoming Organizational Hurdles". 32

Using Confidence Testing in a Physical Diagnosis Course. 38

PATIENT SIMULATORS IN CLINICAL TEACHING

PRECIS. 45

Evaluation of a Self-Instructional Training Program For Simulated Patients. 47

An Instructional Program Using Patient Instructors as Teachers and Evaluators. 53

A Comparison of Instructional Techniques: Internal-External Vs. External Only. 59

ASSESSMENT OF CLINICAL REASONING

PRECIS. 65

The Underlying Structure of Clinical Problem-Solving: Process or Content?. 67

The "Triple Jump" Exercise - A Structured Measure of Problem Solving and Self Directed Learning. 73

Evaluating Analytic Reasoning in the Management of the Multiproblem Patient: A System for Attaining Reliability and Validity. 78

PROFESSIONAL DEVELOPMENT

PRECIS. 85

Assessing the Moral Development of Medical Students: An Empirical Study. 87

Moral Reasoning and Physicians' Decisions in Cases of Critical Illness. 93

Coping and Ego Development in Internship. 99

PERFORMANCE EVALUATION IN GRADUATE MEDICAL EDUCATION

PRECIS. 105

Utah Continuing Studies on Physician and Physician-in-Training Performance. 107

Failure of Examination Answers to Evaluate Actual Practice Patterns by Medical Housestaff in an Outpatient Clinic. 113

Reliability in Follow-up Evaluations of Graduates as Residents. . . .118

The Effects of Clinically Relevant Multiple-Choice Items on the Statistical Discrimination of Physician Clinical Competence. 124

PREDICTION AND SELECTION

PRECIS. 131

The Relationship Between Scores on the New Medical College Admission Test and Performance in the First Year of Medical School: A Preliminary Study. 133

A Factor Analytic Study of the Old and New MCAT Examinations. 139

A Non-Linear Discriminant Analytic Approach to Predicting Medical Student Performance. 145

A Structured Interview for the Selection of Physician's Assistant Students. 151

EVALUATION OF TEACHING

PRECIS. 157

How Medical Students Use Objectives. 159

What Should Students Evaluate? A Comparison of Faculty and Student Perceptions. 165

Teaching Effectiveness: A Comparison of Student Ratings and Graduate Effectiveness. 171

Comparison of Faculty and Chairmen Perceptions About Promotion Criteria. 177

CLINICAL PERFORMANCE RATING SCALES

PRECIS. 183

The Development of Rating Scales to Assess the Clinical Performance of Medical Students. 185

Inter-Rater Reliability of Clinical Performance Ratings. 191

The Relationship Between Reported Activities and Clinical Evaluations of Third-Year Medical Students in a Medicine Clerkship. .197

MEANING, COMMUNICATION, AND MEMORY

PRECIS. 203

The Communication of Uncertainty. 205

Physician-Patient Communication: Interpretations of Non-Technical Phrases. 208

Clinical Experience and the Structure of Memory. 214

ENHANCING CLINICAL TEACHING

PRECIS. 219

Influence of Sequencing in Clinical Tutorials for Second Year Medical Students. 221

Development, Implementation and Evaluation of a Program to Improve Clinical Teaching. 227

Verbal Participation and Outcomes in Medical Education: A Study of Third-year Clinical Discussion Groups. 233

Written vs Oral Feedback: Their Effect on Learning in an Internal Medicine Clerkship. 239

ASSESSMENT IN CONTINUING MEDICAL EDUCATION

PRECIS. 245

Using Audit as a Needs Assessment and Evaluation for a Continuing Education Program. 247

Continuing Medical Education: An Assessment of Impact Upon Neonatal Care in Community Hospitals. 252

Medical Care Evaluation: An Experience in Continuing Medical Education. 258

Development and Evaluation of a Community-Based Pulmonary Education System. 264

Physician Practice Profiles. 270

STUDENT SATISFACTION AND STRESS

PRECIS. 277

Identifying Satisfying and Stressful Events in Medical School. . . . 279

Medical Vs. Graduate Students: Socialization and Stress. 285

Dogmatism and Values Orientation of Residents in Family Practice. . . 291

ISSUES IN TESTING AND GRADING

PRECIS. 297

The Effects of Commercial Tutoring on Medical Licensure Examinations. 299

Curriculum Bias: A Study of a College-Certifying System for Students at Two Basic Medical Science Schools. 305

Medical Students Learning Processes and Their Relation to Test Performance and Attitudes Toward Item Formats. 311

SYMPOSIA

New Experiments to Improve Premedical and Medical Education. 317

Research in Residency Training For Primary Care. 329

United States Citizen Foreign-Trained Medical Students Accepted With Advanced Standing Into American Medical Schools: Background, Reentry and Prognosis. 337

Results of a National Study of a Computer Simulation (CASE) and The American Board of Internal Medicine's Recertification Examination: THE MERIT PROJECT. 345

Perspectives on Humanistic Education. 355

Research on Women Physicians: Recent Empirical Findings and Areas for Future Investigation. 367

Medical Problem Solving: A Reanalysis and Alternative Perspectives. 377

SYMPOSIA, continued

Alternatives to Traditional Methods of Curriculum Evaluation. 389

New Horizons in Cost Containment Curricula. 399

Patient/Health Education: Training for What?. 407

The Clinical Performance of Medical School Graduates:
Theoretical, Methodological, and Empirical Perspectives. 417

The Role of Decision Analysis in Medical Education and Practice. . . 427

INDEX - Authors and Symposia Participants. 439

POSTER SESSIONS

PRECIS

THREE YEARS OF FACULTY DEVELOPMENT PERSPECTIVES OF THE PAST, PRESENT, AND FUTURE

Three years of experience in faculty development will be shared through display and discussion. Handouts will describe present activities and evaluation results.

DEVELOPING A STUDENT DATA BASE FOR INSTITUTIONAL RESEARCH AND EVALUATION

The UNC School of Medicine has developed a computerized student data base for curriculum evaluation and institutional research. The staff of the project wishes to share its experience and its solutions to several general problems attaching to development of systems of this type.

A METHOD FOR PREDICTING NATIONAL BOARD: PART I SCORES

A strategy for predicting student National Board: Part I scores by using a practice examination and linear regression was developed. The strategy was used to provide predictive quantitative information to students to help allay their anxieties toward the actual National Board: Part I examination.

ASSESSMENT OF THE ABILITY TO OBSERVE AFFECTIVE BEHAVIOR IN MEDICAL INTERVIEWS

This paper describes the development of a reliable instrument for the assessment of medical students' observational skills in patient interviews. The final instrument, known as the Observational Skills Assessment (OSA) consists of a 20 item video tape test with an internal consistency reliability of r = 0.80.

Continued. . . .

USE OF MULTIPLE PERSPECTIVES IN DETERMINING CURRICULUM OR TEST CONTENT

This study describes the use of ratings by basic science faculty, clinical faculty and students in reviewing basic science curriculum objectives. Results indicate that: 1) the method works, 2) 30 to 40 percent of basic science objectives represented competencies judged essential to be maintained, 3) there is moderate agreement among ratings by module authors, basic science faculty, clinical faculty and students.

EVALUATION STRATEGIES FOR RESIDENCY TRAINING IN BEHAVIORAL PEDIATRICS, OR, "OVERCOMING ORGANIZATIONAL HURDLES"

Specific organizational features inhibit ideal evaluation of this Behavioral Pediatrics and Developmental Medicine training program for pediatric residents. The poster session will describe the program and evaluation efforts while inviting specific feedback from colleagues regarding alternative evaluative research strategies.

USING CONFIDENCE TESTING IN A PHYSICAL DIAGNOSIS COURSE

Reliability, validity, and acceptability of a novel form of testing were explored. A reliable measure of realism in assessing knowledge mastery was shown to have potential for better understanding certain aspects of clinical judgement.

THREE YEARS OF FACULTY DEVELOPMENT
PERSPECTIVES OF THE PAST, PRESENT, AND FUTURE

Sylvia Shellenberger, Ph.D.*, T. Walter Treadwell, M.D., N. Kay Mashburn, M.Ed.
The University of Mississippi Medical Center
Jackson, Mississippi

Problem Statement and Background

Medical school faculty have spent many years in preparation for the clinical practice of their respective disciplines. However, they have usually had no formal preparation for many other aspects of their roles, such as planning, implementing, or evaluating their instructional efforts. Physician teachers tend to teach in the manner that they themselves were taught. However, in spite of this lack of formal training in "how to teach," one finds large numbers of excellent teaching physicians and scientists among medical school faculty. Significant factors contributing to this undoubtedly include the deep commitment of faculty to help students learn, a fundamental approach of thinking about learning, and the exposure to excellent teacher role models in the education of every physician.

Our Department desired in 1975 to explore the extent to which an increased knowledge of educational concepts and systematic faculty development activities would improve us as teachers and enhance our function broadly in the faculty role. These questions naturally arose: "Can medical educators learn to more effectively teach?", "Are the teaching skills of the educator important in student learning?" and "Is there room for improvement in clinical teaching?" A review of the earlier literature was disappointing in that relatively little educational research had been conducted in medical settings and much research had been criticized by educators as suffering from inadequate design and neglect of theory. [1,2,3,4] Some felt that there were few outcomes of value to the practitioner or for improving teacher-education.[5,6] Also, many conclusions had been based on student rating scales which had both supporters and non-supporters in terms of acceptance of their validity as an instrument for measuring teaching quality. [7,8,9,10,11]

More recent literature has suggested that teacher effectiveness was important to learning and that teacher training of medical school faculty was effective. [12,13,14,15,16,17] Themes evident in the literature are reflected in Gazda's suggestion that "Mastering the academic subject matter is a prerequisite to good teaching but is no guarantee of it." [18] For medical educators, the implication is that being an effective clinician is no guarantee of being an effective clinical teacher. Jason's suggestion that "it is more difficult to teach a discipline that it is to practice it" provides much food for thought for all physician teachers.[19]

---

*Requests for reprints should be sent to Sylvia Shellenberger, Ph.D., Assistant Professor, Department of Family Medicine, 2500 North State Street, Jackson, Mississippi, 39216.

Methodology

The findings of the above research were presented to the Department of Family Medicine faculty with the recommendation that an organized faculty development effort would be a most worthwhile investment for the Department. The faculty endorsed this and time provisions were committed by the Chairman.

A proposal was prepared and submitted to H.E.W. for a special projects grant, "A 40 Week Course to Prepare Primary Care Practitioners to Assume Faculty Roles." (DHEW04-D-001452-01-0) The project was approved and funded for July 1976 through June 1978. The project included a series of "mini-courses" and several workshops on areas of graduate education and administration that highly relate to the student and residency programs and to teaching in general. The unique aspect of the project was that an educator worked with the physician faculty in the preparation and planning of materials and activities for the teaching of the educational concepts. The presentations and activities were then "team-taught" by the educator and the various physician faculty who were participating in the project. The approach of having physicians participate in the teaching of physicians about education insured a high degree of relevance to the faculty's needs and contributed to the high level of acceptance by those participating. The educational concepts were taught within the context of medicine and in such a manner that the skills acquired by the participants could be immediately applied to their academic endeavors. The major curriculum areas addressed were:

Adaptation to The Faculty Role
Effective Communication
Teaching-Learning
Instructional Methodologies and Materials
Curriculum Development
Educational Evaluation
Administration

Activities provided a broad exposure to "education." A positive and measurable impact on the faculty's attitude about education and on their teaching skills was evident. Finding the experience to be highly valuable, they wanted to continue with another project. A new grant application was developed and funding was obtained for support of the activities (DHEW #1-D15-PE-54006-01). The purposes of the present project are: to increase the number of full-time family physician faculty teaching in Family Medicine; to conduct a systematic program of activities to increase the competence of newly-recruited family physician faculty in pedagogical skills and other areas of academic responsibility; and to improve the teaching skills of those already teaching in Family Medicine either at the full-time, part-time or preceptor level.

Specific objectives of the project are:

1. To conduct ten in-state educational workshops for clinical instructor preceptors, two in each of the five state regions.

2. To develop and implement a coordinated educational course for practicing family physicians who have expressed an interest in full-time teaching. The course includes:

   a. Attendance at two regional workshops as described above.
   b. Participation in an educational "correspondence course" including assigned texts and lessons in the form of monographs and lesson tests prepared by the project staff.
   c. Participation in an intensive five day orientation to full-time teaching in Jackson.

3. To implement a Faculty Development program designed for newly-employed faculty physicians and to continue educational training of experienced faculty.

4. To hold a three-day educational workshop designed for Family Medicine residents from throughout the Southeast who have an interest in a teaching career.

5. To implement a teaching component for residents designed to provide an overview of the responsibilities and rewards of full-time teaching and to assist in improving resident teaching skill.

Several advantages of the program as designed appear to enhance its effectiveness. One advantage is that all levels of faculty are reached, including full-time (new and more experienced), part-time, and preceptors. Another advantage is that residents are addressed in a manner which improves their teaching skills and stimulates their interest in a future teaching career. In addition, there is no reduction in faculty income in this program as would be incurred in a one-or two-year clinical traineeship. One further advantage is that along with regularly scheduled classroom activities, physician-participants are observed and critiqued with reinforcement and feedback in the clinical care setting on a continuing basis.

Results

Posters displaying the activities of each component will be presented at the session. So that others may maximally benefit from our experiences to date, evaluation of the various components will be distributed at the poster session in booklet form. From this, participants at the session may better determine the suitability of similar activities for their own respective settings. These evaluations include:

Formative Evaluation--Participants of each component complete an evaluation form on termination of each activity (workshop, rotation, mini-course, 5-day orientation, correspondence course, faculty development session). Results are compiled and used to improve subsequent activities.

Structural Evaluation--Records are kept related to time, date, and place of activities, topics discussed, and those attending.

Assessment of Learning and Attitudes Toward Teaching--Pre- and post-tests are given where appropriate (resident component, new-faculty course, faculty development meetings, 5-day orientation, 3-day educational workshop for residents).

Interviews--At the conclusion of certain components, participants are asked for information regarding effectiveness and relevance of the activities and recommendations for improvement.

Follow-up Survey--Participants will be followed longitudinally to assess their subsequent increased or decreased involvement in medical education.

Discussion

Questions which the presenters will be prepared to discuss and on which they would like feedback include:

1. Methods for determining objectives and content of each component (new-faculty courses, faculty development seminars, resident teaching rotation, workshops, correspondence course, 5-day orientation)
    a. Needs assessment
    b. Personal interview
    c. Survey questionnaire
2. Design of each component
    a. Incorporation of previous materials found to be effective
    b. Review of pertinent literature
    c. Determining topics to be covered, format, methods of teaching, and audio-visual aids
3. Implementation of each component
    a. Scheduling time to meet
    b. Preparing and organizing each session or activity
    c. Obtaining support and assistance from others
4. Evaluating each component
    a. Pre-post assessments
    b. Attitude questionnaires
    c. Interviews
    d. Observation
    e. Videotape
    f. Site visits
    g. Longitudinal follow-up

Conclusions and Implications

Three years of experience in systematically attempting to increase the effectiveness of faculty teaching skills have provided us with perspectives for present and future goals of a faculty development program which we will gladly share and discuss with poster session attendants.

FOOTNOTES

1. Reed, H. B. "Duplications for Science Education of a Teacher Competence Research." Science Education, December, 1962, pp. 473-486.

2. Barr, A. S., et al. "Second Report of the Committee on Criteria of Teacher Effectiveness," Journal of Education Research, 1953, Vol. 46, pp. 641-648.

3. Levin, H. "A New Perspective on Teacher Competence Research," Harvard Educational Review, 1954, Vol. 24, pp. 98-105.

4. Mitzel, H. E. "Teacher Effectiveness," In C. Harris, ed. Encyclopedia of Education Research, Third Ed., New York: The MacMillan Co., 1960.

5. Barr, A. S., Davis, R. A., and Johnson, P. O. Educational Research and Appraisal, Philadelphia: J. B. Lippincott Co.

6. Howsam, R. B. "Teacher Evaluation Facts and Folklore," National Elementary Principal, 1963, Vol. 43, pp. 6-18.

7. Rodin, M. and Rodin, B. "Student Evaluation of Teachers," Science, 1972, Vol. 177, pp. 1164-1166.

8. Bryson, Rebecca. "Teacher Evaluation and Student Learning: A Reexamination," The Journal of Educational Research, September, 1974, Vol. 68, no. 1, pp. 12-14.

9. Frey, Peter. "Student Ratings of Teaching: Validity of Several Rating Factors," Science, October 5, 1973, Vol. 18, pp. 83-85.

10. Rippey, Robert. "Student Evaluation of Professors: Are They of Value?" Journal of Medical Education, October, 1975, Vol. 50, no. 10, pp. 951-958.

11. Doyle, K. and Whiteley, S. "Reply to Rodin and Rodin," Measurement Services Center, University of Minnesota, November, 1972.

12. Centra, J. A. "Effectiveness of Student Feedback in Modifying College Instruction," Journal of Educational Psychology, 1973, Vol. 65, no. 3, pp. 395-401.

13. Pambookian, H. S. "Initial Level of Student Evaluation of Instruction as a Source of Influence on Instructor Change After Feedback," Journal of Educational Psychology, 1974, Vol. 66, pp. 52-56.

14. Rous, S. N., Banford, Gromash. "The Improvement of Faculty Teaching Through Evaluation: A Preliminary Report," Journal of Surgical Research, 1971, Vol. 11, pp. 311-315.

15. Rous, S. N., Banford, Gromash. "The Improvement of Faculty Teaching Through Evaluation: A Follow-Up Report," Journal of Surgical Research, 1972, Vol. 13, pp. 262-266.

16. Pochyly, Donald. "Problem-Oriented Faculty Development in A Medical School." Educational Horizons. 1976-77, Vol. 55, no. 2, pp. 92-96.

17. Snope, Frank C., et al. "Teacher Training in Family Practice: A Preliminary Report." The Journal of Family Practice, 1977, Vol. 5, no. 1, pp. 134-135.

18. Gazda, G. M. "Systematic Human Relations Training in Teacher Preparation and In-Service Education," Journal of Research and Development in Education, 1971, Vol. 4, no. 2, pp. 47-51

19. Jason, Hilliard. Personnel Communication. Consultative visit to Dept. of Family Medicine, University of Mississippi Medical Center. 1975.

Developing a Student Data Base for Institutional Research & Evaluation

Claudia M. Cheatham
Carol Q. Porter
Charles P. Friedman
University of North Carolina at Chapel Hill*

I. Description of the UNC Student Data Base

Overview

In the Spring of 1978 the Office of Medical Studies of the University of North Carolina School of Medicine initiated development of a longitudinal student data base and a mechanism for its use. During the first developmental year, the staff has dealt with a number of issues in the areas of software design, data collection and management, the legalities surrounding confidentiality, modes of utilization, and related matters which we feel would be useful to share with others interested in developing similar systems.

Above all, the data base represents a commitment on the part of the UNC School of Medicine to a standing resource for institutional research and curriculum evaluation. Although the data base shares many properties with computerized record keeping systems and data files to inform longitudinal research studies, its detailed structure and projected patterns of utilization differ significantly. Unlike longitudinal research projects, the data base has no termination date and no *a priori* investigative questions guiding its operation--although it can be used for longitudinal research. Unlike record-keeping systems, the UNC data base is not designed to execute any particular administrative task--although it can be helpful to such tasks.

Staff of the project do not collect information; rather they recast information generated as part of the ongoing educational operations of the school. While the staff of the project will be undertaking some studies on their own initiative, they prefer to operate in a responsive posture. This implies letting others formulate the questions and then, through a process of consultation and negotiation, to use the data base to address the questions.

To our knowledge, few other medical schools possess the type of student data base developed at UNC. A survey conducted in the Spring of 1979 by Richard Randlett of AAMC reveals that six schools have similar systems and we are aware of two more such programs not listed in the AAMC survey.

The data base is currently staffed by a faculty member with professional interest in research and evaluation, a project coordinator and a systems programmer. It operates under the guidance of a steering committee comprising nine medical school faculty and administrators. During the early months of development the steering committee and project staff addressed the potential problems attaching to day-to-day operation of the base, most complex and time-consuming of these being the issues of confidentiality and access. With the help of university attorneys, a formal policy was established to restrict access to the system. This document also defined types of information that may be released from the data base.

---

*Reprints available from: The Office of Medical Studies, University of North Carolina School of Medicine, Chapel Hill, North Carolina 27514

Why a Data Base?

Questions concerning the predictors of success in medical school, the performance of medical students, the efficacy of basic science and clinical education, and post-graduate career patterns of medical students arise frequently in the ongoing administration of a medical school. Such questions are posed by faculty, administrators, and students alike. The existence of comprehensive, accurate data files, efficient access to these files, and capability for economical data analysis facilitate these types of institutional and evaluative research. A computerized data base is also the most economical way to deal with the storage and management problems of massive amounts of information. During a time of increased emphasis on evaluation and accountability by public and private funding sources, a computerized system is a logical solution to responding to repeated requests for information and statistical analyses.

Contents of the Data Base

After thorough deliberations with the steering committee, a review of existing data and an assessment of future needs, the following data were included in the computerized file for all UNC medical students beginning with the class that entered in 1970:

1) Admissions data (primarily from data provided by AMCAS and responses to a biographical questionnaire completed by each entering student).
2) A "control file" (a trajectory of each student's progress through the curriculum).
3) Records of all final examination scores (including National Board and locally-prepared exams), comprehensive year-end grades, and clinical clerkship and electives grades.
4) Location of clinical clerkships (Many UNC students take all or significant parts of their clerkships outside of the Chapel Hill campus).
5) A list of electives taken.
6) Responses to a questionnaire administered upon graduation from medical school.
7) Internship, residency and practice type and location.

This information base is seen as providing information necessary to answer the great majority of questions that may arise. In addition, this data "core" can be merged temporarily with files of information required for special studies employing data not routinely stored.

Confidentiality and Security

Individuals are increasingly aware of intrusions into their personal lives as a result of information gathering from various sources, including universities. Being acutely conscious of the issues of privacy and confidentiality, as well as the dictates of the Family Educational Rights & Privacy Act, the data base staff and the steering committee established stringent security measures to ensure appropriate use of the information. To these ends, the following policies and features were incorporated:

1) All data about students are stored according to a scrambled identification number, the key to which is known only to project staff.
2) Each data file in the system is protected by a unique password known only to the staff. The data file cannot be read or modified without this password.
3) Only statistical analyses of anonymous raw data are available from the data base. Even aggregate data may not be released if the group studied is sufficiently small or unique to reveal the identity of individual students comprising the group.
4) Only full-time faculty, administrators and students at the UNC School of Medicine may make requests for information. Each request is reviewed by the staff under guidelines established by the steering committee. Any request not considered to fall strictly within the boundaries of the guidelines is referred to the Associate Dean of Student Affairs and two members of the steering committee, who then decide whether the request can be honored.

Operating Procedure for the Data Base

As the data base enters its second year, routine operating procedures are beginning to take shape. These procedures have evolved from experiences handling requests during the past 14 months, and the relative efficiency of various methods. At present, we consider the following to be a reasonable and workable modus operandi.

A request to the coordinator of the data system is entered onto a standardized form. Included on the form are the name and departmental affiliation of the requestor, the nature of the request (including for what purpose will the analyses be used, what variables specifically are to be scrutinized, what questions are to be answered), and the requested date of completion of analysis.

After a request is entered, the staff of the data base take the role of consultant. Often a requestor does not know which particular data or statistical procedures are appropriate to answer the questions posed, making it necessary for the staff to provide guidance in these areas. Sometimes the requestor wants a specific problem investigated but requires assistance in articulating the problem in the form of questions consistent with statistical research. For many of these same reasons, interpretation of the results is provided to the requestor upon completion of the study. Current policy is to honor requests on a "first come, first served" basis, a system which sometimes conflicts with requestor deadlines and priorities. We offer as much flexibility as we can in this regard, but we recognize that this may be a source of problems in the future as the number of requests increases.

At present, funds for the student data base are provided through the operating budget of the Office of Medical Studies. There is no charge to the requestor for analyses of data in the base. When a request is made which requires data not stored in the base, the requestor is expected to provide the additional data in computer-readable form so that only simple programming is required to temporarily merge the files.

System Design Objectives & Software

System design objectives of the longitudinal student data base were to allow maximum flexibility in updating and retrieving data, and to make efficient use of computer disk storage space. For these reasons the following were important considerations in designing the system:

1) Since data items collected about medical students change from year to year as instruments change, the system should allow introduction of new variables without requiring updating of unaffected records. Conversely, once a data item is deleted, disk space should not be wasted for the obsolescent variable in all future records.
2) The system should recognize that many academic careers deviate from the standard of four consecutive years. Because the population of students required for a particular analysis is dependent on the question being addressed, the system must be able to economically assemble working files comprising data on the set of individuals appropriate to address particular questions.
3) The UNC curriculum consists of four years. A group of students progressing through a given year of the curriculum may be considered as a distinct cohort. This feature suggests treatment of each such cohort in a given academic year as a logical unit of data record.
4) Software for a longitudinal student data base must provide data management capabilities, statistical analysis procedures and report writing features. Data in the system should be readable by a variety of computer programs and statistical packages.

For these reasons, the UNC student data project elected to use the Statistical Analysis System (SAS)* as software for its data base, storing data on off-line disk with tape backup. SAS is an integrated system which provides data management capabilities far superior to similar systems such as SPSS. Using a language resembling PL/1, the user can create, transform or delete variables, and datasets can be easily created, merged, concatenated and updated. SAS will accept system files created by other statistical packages and SAS files can be converted to card-image files readable by other programs. A large number of statistical procedures are provided.

SAS has proved to be an economical system. It requires a relatively small user region so batch programs run quickly, and most importantly, its simplified command language has eliminated the need for extensive programming. SAS is available to the data base at no cost beyond computer time, since it is maintained by the university's computation center. Interactive access to the system would be possible should the need arise.

System Design

The data storage system consists of eight information files, each containing data collected during one of eight time intervals: pre-admission, first year, second year, third year, fourth year, internship, residency, and practice. Each of these files is in turn divided into subfiles for each academic year of interest to our project: 1970-71, 1971-72, etc. Each item of information about each student is then stored in the appropriate academic year subfile of the appropriate information file. For example, the first year exam scores of all students who took the first curriculum year 1975-76 are stored in the "1975-76" subfile of the "first year" information file. Pre-admission data is keyed to the year in which a student entered medical school, and post-graduate data is keyed to the academic year in which the student graduated. If a student repeats, say, the first year, he/she will appear in two subfiles of the first year file. Any combination of these subfiles can be merged as required for a particular analysis.

*SAS Programmer's Guide, Anthony J. Barr, et al, SAS Institute, Sparks Press, Raleigh, NC, 1977.

A control file "remembers" the academic years in which each student undertook each year of the medical school curriculum. This control file provides the system with pointers to the subfiles required for any analysis. The control file subfile design resolves problems associated with storing and retrieving data on students whose academic career deviates from the standard of four consecutive years.

II. Representative List of Questions Presenters will be Prepared to Discuss

1. What are the implications of the Family Educational Rights & Privacy Act?
2. How has data been used so far (examples of requests)?
3. For what new questions might the data base be used?
4. How does the system interact with record-keeping in the Registrar's Office?
5. How can the system be modified to suit other needs, such as scheduling?
6. What is the best way to "publicize" the existence of the system and the uses to which it may be put?
7. How can responsible use of the data be encouraged?
8. How can student concerns about confidentiality be satisfied?
9. How can the files be manipulated for best maximum efficiency?
10. How is interaction with computer support services, such as keypunching best established?
11. When and how are priorities in handling requests determined?
12. What routine services can a computerized data system perform to enhance the daily operation of administration of the medical school?
13. Data collection problems: How can information be made comparable over years when data items change with new instruments?
14. What is the most effective way to instill "computer consciousness" in the designing of forms from which data will be transferred to the data system?
15. How can skeptics be convinced of the usefulness, efficiency, confidentiality of a computerized system of data?
16. Cost: What have we spent? Who pays?
17. How does one implementing such a system deal with "computer phobia"?

III. Description of Poster Session Display

The UNC data project lends itself well to graphic representations. We foresee a graphic display divided into three separate sections:

1. Handouts and posters illuminating considerations prior to implementation of a system.

   Evaluating of the need for a system
   Defining the purpose of the system
   Formalizing security and confidentiality policies
   Evaluating staffing needs
   Determing administrative locus of system
   Selecting suitable computer software
   Formulating operational procedures and policies

2. A chart giving graphical representation of computer system design, including both the architecture and the content of the system.

3. An example of a typical use of the system traced from definition of problem through summarized results of analyses. This display would include a specimen request form for a hypothetical study, computer printout and interpretive report of the results.

A METHOD FOR PREDICTING NATIONAL BOARD: PART I SCORES

William E. Sorlie, Diane Essex, Ph.D.

University of Illinois-Urbana

Individuals in medical education have long been concerned with student success in medical school. In fact, one of the first studies appears in the literature in the mid 1920's (1). Since that time, numerous investigations have been completed and published; the focus of much of this research has been on the prediction of success in the first two years of medical school (2). Researchers have used the prediction model to select from among all applicants those most likely to succeed in medical school (2). Other investigators have attempted to predict success of already admitted students (3). Even though the goals of selection procedures and prediction of success post admission differ somewhat, there are shared similarities. Among these are the variables utilized. Consideration is typically given to premedical achievement, often as reflected in undergraduate grade point average (4), and scientific aptitude, generally as measured by the Medical College Admission Test -- MCAT (2). Success in medical school (the criterion) has been defined somewhat differently by various researchers; for example, success has been operationalized as first-year medical school grades (3), performance on a year-end certifying examination (5), or rank in the first-year class.

As has been pointed out previously, the results from past prediction research have been somewhat disappointing (4). Despite the continued search for correlates of success in medical school, validity coefficients have remained relatively low. It is seldom the case that the validity coefficients exceed 0.35 (4,6,7). Given these results, one may wonder about the profitability of further attempts to predict the success of already admitted students. Particularly, one must certainly question the continued use of premedical achievement data and aptitude test scores (e.g., MCAT scores).

If the work in the area of prediction were to continue, investigators may want to redirect their efforts to the early detection of potential failures after students have become involved in the educational program. One remarkable difference in these redirected efforts, as compared with previous efforts, would be the selection of predictor variables. Another difference would be the purpose of such efforts; the goal would be instructional intervention in which students identified as "high risk" would be counseled and would receive instructional remediation tailored to their specific weaknesses.

A first step in the redirection of prediction research has already been taken (4). In the first phase of a study by Thomson, Essex, and Sorlie (4), premedical achievement variables were utilized to predict performance on two certifying examinations, one of which was the National Board: Part I. In phase two of the study, these predictions were refined when performance on faculty-developed basic science examinations were incorporated into the predictions. Since the goal was early identification of potential academic problems, data

---

Reprints: William E. Sorlie, Schools of Basic Medical Science and Clinical Medicine, University of Illinois, 190 Medical Sciences Building, Urbana, IL 61801.

were gathered in the first three months of the educational program. Validity coefficients with performance on basic science examinations as the predictors approached 0.70.

A second step in these new efforts would be to search for predictor variables that are more relevant to the criterion under consideration. The findings of such an effort are presented in this paper. An attempt was made to determine if performance on a practice board examination (PreTest) had value as a predictor of performance on the actual National Board: Part I.

In many medical schools, the National Board examination is used as a certifying instrument; consequently, students tend to be extremely anxious about the examination during the several weeks preceding its administration. Contributing to this anxiety is the fact that formal instruction in some of the basic medical sciences may be as much as 12 months removed and in-depth review is essential to success on the National Board. Unfortunately, the anxiety and need to review extensively cause many students to "drop out" of the regularly scheduled lectures, seminars, laboratories, and demonstrations which are presented during the several weeks prior to the examination. From an instructional and educational standpoint, one must certainly be concerned about the amount of material students miss during this "drop out" period. Consequently, information relevant to the student's predicted National Board performance could allay his/her anxieties and encourage continued participation in the on-going educational program.

Background

The School of Basic Medical Sciences at Urbana-Champaign (SBMS-UC) has been involved in an innovative educational approach to the preclinical year(s) of medical education. When SBMS-UC opened in the Fall of 1971, basic science faculty and local practicing physicians designed a written curriculum to provide medical students with the skills and knowledge necessary to allow entry into a clinical training program after about 10 months of study. The written curriculum contains 13 basic medical science disciplines and is organized into a five volume document. The curriculum contains over 400 self-instructional units and incorporates over 4,000 specific learning objectives. Each curriculum unit includes all necessary learning components (objectives, key words, reading assignments, self-tests) to allow students to proceed individually (5).

In addition, to demonstrate the relevance of basic science to clinical medicine, each of the 400 learning units is related to 1 of 9 clinical problems or diseases. While studying the learning units, each student is provided multiple opportunities to interact with local practicing physicians and their patients which serve to correlate basic science content and clinical problems or diseases. The specific patient problem is used as the entry point into interdisciplinary principles of basic science germane to the understanding of medicine as a whole. Student progress is assessed by school diagnostic examinations, a college certifying examination, and the National Board: Part I. In April of each year, a practice examination (PreTest) is administered to students.

PreTest Description

The National Board: Part I PreTest is a commercially developed test that is available for a fee from its publisher and is designed to parallel closely the actual Part I examination. The disciplines tested, item type, time limits, and scoring procedures are similar to the actual Part I examination. However, the PreTest is half the length of the Part I test.

Students record their responses on mark-sense sheets; these sheets are then returned to the PreTest Service where they are scored. Scores, adjusted to the length of the Part I examination, as well as detailed references for each item, are returned to students several weeks prior to the administration of the actual National Board: Part I examination.

METHOD

Subjects. Data were gathered from basic medical science students in three different classes. For the years 1976, 1977, and 1978, the PreTest total score (adjusted to the length of the Part I examination) and the National Board: Part I total score were assembled for each student: $n_{76} = 58$, $n_{77} = 64$, and $n_{78} = 65$.

Procedure. The PreTest was administered in early April; the Part I examination was administered in mid-June of each year. Further, at the end of each academic year, students were asked to complete a questionnaire in which they were given an opportunity to evaluate various aspects of their first-year experience. The questionnaire consisted of about 50 items, some of which have multiple parts. Items requiring responses to structured questions, as well as open-ended questions, were included in the questionnaire.

Within the questionnaire, several different curricular and instructional areas are covered. Of particular interest in this study were student perceptions about the value of the practice Part I examination (PreTest) and the quantitative information provided relative to their estimated National Board: Part I scores.

Analyses. Relative to the questionnaire data, basic descriptive statistics, i.e., frequencies, means, and standard deviations were computed. These statistics were then examined to determine if students found the PreTest and associated quantitative data information useful in reducing their anxiety relative to the Part I examination.

Further, linear regression was used to determine the extent to which the PreTest total score could be used to predict National Board: Part I total scores. Additionally, in order to examine the stability of the mutliple correlation, cross-validation was performed. For example, the regression equation developed on the 1976 student group (screening sample) was then applied to the 1977 group (validating sample). The equation based on the 1977 group (screening sample) was then applied to the 1976, as well as 1978, student group (validating samples). The multiple correlations were then examined for shrinkage (8).

RESULTS

Displayed in Table 1 are the results of the regression analyses for 1976, 1977, and 1978. The PreTest mean and standard deviation are displayed in column two; the regression equation ($\hat{Y} = bX + a$) is given in column three; the multiple correlation (R) appears in column four; descriptive information for the National Board: Part I total scores appears in column five (predicted mean total score) and column six (obtained mean total score).

It can be seen from the table that the PreTest mean (column two) ranged from 541.03 in 1977 to 480.53 in 1976. From column three, it can be noted that the regression coefficient ($\beta = 1.21$) and intercept ($\alpha = -169.17$) calculated from the 1977 data differed substantially from those obtained in 1976 and 1978. Construction of the 95% confidence interval about the regression coefficient and intercept from the 1977 analysis confirmed that they differed significantly from their 1976 and 1978 counterparts ($1.02 < \alpha < 1.40$; $-274.35 < \beta < -63.99$). However, from column four, it can be observed that the multiple correlation was rather stable over the three-year period; it ranged from 0.87 in 1976 to 0.79 in 1978.

TABLE 1

Summary of Regression Analyses for 1976, 1977, and 1978

Screening Sample	Pre-Test Total Score ($\bar{X}$)	Regression Equation	Multiple Correlation	National Board: Part I Predicted ($\hat{\bar{Y}}$)	Total Scores Obtained ($\bar{Y}$)
1976, n=58	$\bar{X}$=480.53 s=86.00	0.96X+ 11.94	0.87	$\hat{\bar{Y}}$=471.04 s=81.94	$\bar{Y}$=471.03 s=93.99
1977, n=64	$\bar{X}$=541.03 s=70.17	1.21X-169.17	0.85	$\hat{\bar{Y}}$=486.34 s=89.14	$\bar{Y}$=486.33 s=100.34
1978, n=65	$\bar{X}$=540.43 s=67.67	0.94X-15.14	0.79	$\hat{\bar{Y}}$=492.93 s=63.62	$\bar{Y}$=492.92 s=80.95

From the results of the cross-validation (Table 2), it was determined that the multiple correlations remained extremely stable. For example, the multiple correlation calculated from the 1976 data was 0.87; when the 1976 regression equation was applied to the 1977 data, the multiple correlation was 0.85. Thus, it appears that the magnitude of the multiple correlations was not due to the operation of chance factors specific to any one of the (screening) samples.

TABLE 2

Results of the Cross-Validations Based on Regression Equations Obtained in 1976, 1977, and 1978

Screening Sample	Validating Sample	Multiple Correlation
1976	1977	0.85
1977	1976	0.87
	1978	0.79
1978	1977	0.85

Additionally, on the basis of responses to relevant items on the student questionnaire, it was noted that about 60% of the students reported that the Pre-Test and National Board: Part I contained comparable material. From examination of responses to an open-ended question, it was observed that students reported they found the practice examination useful in terms of a sample of what might appear on the actual Part I examination; they also reported that the quantitative information had alleviated their anxieties and allowed them to be more efficient in their review. Responses to the relevant questions on the student questionnaire were consistent across all three classes of basic medical science students.

Summary

Data were gathered from students in three different basic medical science classes. The performance of these students on a practice National Board: Part I (PreTest) examination and on the actual Part I examination was studied. It was found that the Pre-Test total score was an excellent predictor of the National Board: Part I total score. Using the results of the 1978 data as an example, the School was able to predict, for 70% of the students, their actual Part I scores ± one standard error of estimate (50 points). The prediction success increased to almost 84% for ± 75 points. Less than 5% of the predicted scores did not fall within the ± 100 point range. Under cross-validation, it was observed that the multiple correlations were extremely stable. Further, students reported that the practice Board examination and quantitative predictive information had been useful to them in their preparation or the actual Part I examination. Results were consistent across the three separate classes of basic medical science students.

Although the multiple correlations between the practice examination total scores and actual Part I scores were substantial (e.g., 0.85), students and their academic advisors should be cautioned not to over-interpret the student's predicted National Board: Part I score. Also, it is suggested that students should not use their predicted scores as the sole criterion for determining the optimal time to sit for the National Board: Part I examination.

REFERENCES

1. Bott, E. A. The Predictive Value of College Marks in Medical School. *J. Educ. Res.*, *12*:215-227, 1925.

2. Gough, H. G. Some Predictive Implications of Premedical Scientific Competence and Preferences. *J. Med. Educ.*, *53*:291-300, 1978.

3. Gough, H. G., Hall, Wallace B., and Harris, R. E. Admissions Procedures as Forecasters of Perfomance in Medical Training. *J. Med. Educ.*, *38*:983-998, 1963.

4. Thomson, D. S., Essex, D .L., and Sorlie, W. E. A Two-Phase Strategy for the Early Identification of Potential Academic Problems in Medical School. *Proceedings of the Seventeenth Annual Conference on Research in Medical Education*, 1978, pp. 3-8.

5. Sorlie, W. E., Anderson, J. D., Gamble, T. E., and Bloomfield, D. K., "A One-Year Program in Basic Medical Sciences." *J. Med. Educ.*, *48*:371-373, 1973.

6. Richards, J. H., Jr., and Taylor, C. W. Predicting Academic Achievement in a College of Medicine from Grades, Test Scores, Interviews, and Ratings. *Educ. Psych. Msmt.*, *21*:987-994, 1961.

7. Crowder, D. G. Prediction of First-Year Grades in a Medical College. *Educ. Psych. Msmt.*, *19*:637-639, 1959.

8. Kerlinger, F. N. and Pedhazur, E. I. *Multiple Regression in Behavioral Research*. New York: Holt, Rinehart, and Winston, Inc., 1973.

ASSESSMENT OF THE ABILITY TO OBSERVE AFFECTIVE BEHAVIOR IN MEDICAL INTERVIEWS

Robert B. Cowan, Jr., M.D., Ronald A. Krol, Ed.D., George M. Curtis, M.D.

The effective teaching of medical interviewing and its evaluation are of prominent concern as consumers demand more sensitivity from their physicians and as the medical profession realizes the role of the physician's psychosocial concern in patient compliance and in the incidence of malpractice claims. Certainly, the psychological distress which is highly prevalent in medical patients and physicians' widely varying ability to identify and intervene with that distress, are equally compelling reasons for improving the teaching of medical interviewing which attends to the feelings of distress. In this developing area, relatively few researchers have focused on specific issues. Three factors have substantial impact on the interview process; the interviewer's ability to observe affective behavior, the interviewer's personality and consequent attention to his/her own feelings, and finally, the interviewer's own affective and cognitive behavior in the interview. (1,2)

The interviewer's ability to infer feelings from verbal and nonverbal affective behavior of the interviewee is a critical first step in the sequence that also includes clarifying those affects, integrating them into an overview of the situation and making appropriate interventions. The area of observation of affective behavior is a skill which lends itself to evaluation as has most recently been demonstrated by Werner, Kagan, and Schneider, with Form D of the Affective Sensitivity Scale (A.S.S.). (3)

The Observational Skills Assessment (O.S.A.) was developed specifically for evaluation of the ability in medical students and primary care residents to observe affective behavior in medical patients. For this purpose the O.S.A. appears to have several advantages over the A.S.S., all the interview situations are medical with a single focus of inquiry, "how is the patient feeling?" The O.S.A. is easy to administer as it requires only the video cassette, cassette players, score sheet, pencil and approximately 40 minutes of time.

METHODS The format of the O.S.A. on the video cassette consists of a general introduction and statement of purpose, directions, examples, and a context for each interview. Then segments of each interview that contain prominent non verbal affective cues are included as vignettes varying in length from 15 to 90 seconds. After each vignette, three statements remain on the screen for 25 seconds. The subjects are given the direction, "Please select the response (a, b, or c) that best reflects the patient's feelings (or verbalization about his/her feelings) at the end of the vignette." The correct response was what the patient actually said in the interview following the video cut. Incorrect responses are either statements the patient made at other points during the interview or statements which minimize affect such as "I just accepted it and went on." and "Sometimes I feel guilty about it, though." (3)

---

* Requests for reprints should be sent to Robert B. Cowan, Jr., M.D., Department of Psychiatry and Behavioral Sciences, University of Texas Medical Branch, Galveston, Texas 77550. We would like to thank John Schneider for his valuable advice, Charles Eddlemon for the television production, and Wendy Tomasko and Rose Farkas for their invaluable clerical help. We would like to acknowledge partial support from the following sources: The Lethe Kalmbach Fund, Consultation Liaison Training Grant through the National Institute of Mental Health, Grant #MH14819-03.

The O.S.A. focuses on the identification of negative or dysphoric affects of anger, sadness, and anxiety, in medical patients with chronic illness (diabetes mellitus), life threatening illness (cancer and cardiovascular disease), depression, substance abuse, and chronic pelvic pain. The interviews are with actual patients with the exception of one spouse. There are 6 women and 5 men; 1 black and 10 whites, with ages 25 to 70 years. The interviews are conducted by an oncologist, a nurse clinician, and by consultation liaison psychiatrists.

Subjects were obtained by incorporating the O.S.A. into several psychiatric clerkships for third and fourth year medical students and the interviewing course for first year medical students so that the affected students would respond in a positive manner to the task they were asked to voluntarily complete. Ninety-three percent complied with the request. Psychiatry residents were also asked to be voluntary subjects.

From nine interviews, 88 items were constructed consisting of a vignette and three statements of feelings for each item. A panel of nine expert raters comprised of psychiatrists and psychologists from two medical schools were selected for their recognized expertise in psychological work, teaching of medical students and research in the area. The experts were asked to take the O.S.A. and to make comments. Their task was viewed as one in which a high rate of agreement of the correct response was interpreted as agreement that the inclusions in the vignette were congruent with the correct response and that the alternative incorrect responses were not overly deceptive. Agreement at the level of 70% or greater was achieved for 40 items. Using these, plus some marginal items two 23 item forms (A and B) were constructed for trial use in a pre-post design.

Because the major goal of this project was to develop a valid and reliable procedure for the assessment of observational skills the following strategies were chosen to accomplish that goal:

a) To determine the equivalency of Form A and Form B, it was decided to administer both forms to a rotation of students on a Psychiatry clerkship. For the forms to be considered equivalent this group should receive the same mean and standard deviation on both forms.

b) To determine if this instrument was able to discriminate by level of training, it was decided to administer the O.S.A. to as many levels of students, residents, and staff as feasible. It was felt that, if this technique did not have a ceiling effect respondents, with more training, would score higher.

c) To determine if the present psychiatry clerkships were teaching any of the concepts measured by this technique. Form A was given as a pre-test at the beginning of two clerkships, and Form B at the conclusion of the same clerkships. It was hypothesized that there would be no change in performance on the O.S.A. since the concepts measured by the instrument are not specifically taught in the present psychiatry clerkships.

d) To determine the convergent validity of the O.S.A., it was decided to administer the O.S.A. and the Affective Sensitivity Scale (A.S.S.), to the same group of medical students. It was expected that there would be a relatively high correlation between these two instruments.

e) To determine the divergent validity of the O.S.A., it was decided to correlate the M.C.A.T. scores for all medical students who had taken the O.S.A. with those students scores on the M.C.A.T. It was expected that there would be a low correlation between these two measures since the M.C.A.T. is believe to be a test of knowledge while the O.S.A. was intended to measure non-cognitive skills.

f) Finally, it was hoped that the internal consistency reliability of the O.S.A. could be increased to the r = 0.70 to r = 0.80 range by an item to total score correlation designed to select items that correlated well with one another. (4)

RESULTS At this stage of development of the O.S.A., data has been gathered from students, residents, and expert raters. These data include medical students in pre-clinical, clinical years, psychiatry residents in various levels of training, and psychiatry faculty at two medical schools. Table I is a summary of the data collected and the mean percentages achieved by the various groups.

TABLE I

Means (in percent) by Level of Training
and Test Form (A or B) for all Respondants

	FORM A			FORM B		
Level of training	N	Mean	Standard Deviation	N	Mean	Standard Deviation
Medical Students (Preclinical)	31	70.0	18.0	31	75.0	13.0
Medical Students (Clinical)	66	75.0	13.2	54	75.0	13.5
Psychiatry Residents	10	71.0	30.0	--	----	----
Expert Faculty Raters	9	88.0	6.2	9	93.0	4.9
Total	116			94		

It should be pointed out that these data include all data collected on questions, on each form, that were common to all groups. Because the O.S.A. is still being developed it includes data before questions were subjected to a detailed item analysis. The data on Form A includes 15 of the 23 questions which were common to all groups, while the data on Form B include 11 of the 23 questions which were common to all groups. It should be noted that the total number of subjects is 116 who took Form A. The 94 subjects who took Form B are the same subjects who took Form A minus the 22 subjects who failed to complete Form B for various reasons.

In order to determine the psychometric equivalency of Form A and Form B, both forms were given to ten third year medical students in a single administration.

In order for these tests to be considered to be equivalent, it would be necessary that they have similar means, standard deviation, and internal consistency reliabilities. Table 2 is a summary of those results.

TABLE 2

Means, Standard Deviations, and Reliability in Form A and Form B

FORM A			FORM B		
Mean	Standard Deviation	Internal Consistency Reliability	Mean	Standard Deviation	Internal Consistency Reliability
16.89	1.83	0.23	18.67	1.94	0.11

Table 2 indicates that the two forms are not equivalent since their means are not equal, although these standard deviations are in the same range. Another problem is the extremely low reliability of both of these instruments. It is well to point out that their reliabilities reflect the lack of a critical item analysis at this stage of the project. This will be discussed later in the paper.

To determine if the present psychiatry clerkships are teaching any of the concepts measured by the O.S.A., it was decided to adminster Form A at the beginning of each rotation to all (N=22) medical students. Table 3 is a summary of those results.

TABLE 3

PRE/POST Means on Form A and Form B for two Psychiatry Clerkship Rotations

Pre-Test (Form A)			Post-Test (Form B)		
N	Mean	Standard Deviation	N	Mean	Standard Deviation
22	16/67	1.75	22	18.67	2.74

From the above data, it appears that there was a gain of two questions. However, if these data are compared to the data in Table 2, it is evident that the students who took Form A and Form B at the same sitting have means which were almost identical to those of the medical students who had taken Form A and then Form B after four weeks in a psychiatry clerkship . These results indicate that these two groups could be combined prior to an item analysis. This was done to increase the N prior to calculating an item to total correlation. In addition, their results indicate that the present psychiatry clerkships do not teach the concepts measured by the O.S.A.

To determine if the O.S.A. is able to discriminate by the number of years of training, various groups of medical students, residents, and faculty were compared. The results in Table 1 indicate that the O.S.A. seems to discriminate medical students by level of training on Form A. For example, on Form A, medical students in their clinical years score higher than medical students in their preclinical years. However, psychiatry residents scores only slightly higher than preclinical medical students (71% vs. 70%) and lower than medical students in their clinical years (71% vs. 75%) on Form A. The same comparisons cannot be made on Form B because similar data on Form B is not available. Clearly, more must be done to determine how scores on the O.S.A. relate to level of training but this cannot be accomplished until a reliable instrument is completely developed.

To select items that would contribute to a reliable instrument, scores on each item were correlated with total test scores (4) on a combined 46 item test for 32 medical students in the clinical years. The internal consistency reliability was calculated to be r = 0.50. Items were eliminated that had the lowest item to total correlation until the 20 items which remained had an internal consistency r = 0.80. This process demonstrated that if the items that were not highly correlated with each other were removed, that a reliable instrument for the assessment of the observational skills in medical students could be constructed.

Furthermore, the item selectec by this process have a high congruence with those selected by the expert panel of raters. That is, all the items selected were also items that have a 70% or greater agreement by the expert panel.

DISCUSSION

The development of the O.S.A. is traced including a premature attempt to develop equivalent/parallel forms. Item to total correlations demonstrate that if the best 20 items are selected for a Form C, that a reliability of 0.80 will be achieved. Face validity is established through the use of the patient's actual feeling statements as correct responses and expert rater agreement at the level of 70% or greater. Discriminant validity is established in a preliminary way, based on performance by level of training on the common items from Form A. However, the psychiatry resident's performance was below that of medical student's, although the numbers of residents were relatively small and they had a much greater variance than other groups. A possible explanation of their performance is that some of them were attempting to select responses based on assumed unconscious processes rather than on observable affective behavior. Werner, Kagan and Schneider had similar findings for psychiatric residents (3).

Convergent and divergent validity remain to be established through the analysis of data currently being collected on the A.S.S and M.C.A.T. Attitudes toward psychiatry and psychological medicine are expected to condry with performace on the O.S.A. since those attitudes covary with physicians' abiiity to identify psychiatric illness (2). Analysis of the data from incoming house officers using O.S.A. Form C and an attitude scale (5), is expected to address this issue.

Much work remains to be done. The correlation of O.S.A. performance with medical students' actual rated interviews would answer questions about convergent and predictive validity of the O.S.A. Furthermore, the correlation of personality factors with O.S.A. performance would be helpful in establishing the degree to which those factors account for O.S.A. score variance.

1) Werner, A. and Schneider, J.: Teaching Medical Students Interactional Skills: A Research-Based Course in the Doctor-Patient Relationship. New England Journal of Medicine, 1974, 290:1232-1237.

2) Marks, J.N., Goldberg, D.P., and Hillier, V.F.: Determinants of the Ability of General Practitioners to Detect Psychiatric Illness. Psychological Medicine, 1979, 9:1-17.

3) Werner, D., Kagan, N., and Schneider, J.: The Measurement of Affective Sensitivity: The Development of an Instrument. Proceedings RIME Conference, 1977.

4) Nunnally, Jum C.: Psychometric Theory, McGraw Hill Book Company, 1978, New York.

5) Krol, R.A., and Curtis G.C.: An Inventory of Attitudes Toward Psychiatry anf Psychological Medicine, paper presented at American Psychosomatic Society, Dallas, Texas, March, 1979.

Use of Multiple Perspectives in Determining Curriculum or Test Content[1]

Reed G. Williams, Southern Illinois University School of Medicine
John H. Shatzer, University of Illinois School of Basic Medical Sciences
Janet McDole, St. Louis Metropolitan Police Department

Faculty are continuously faced with the task of selecting topics for inclusion in a time limited course of study. The responsibility for making these decisions usually is left to individuals assigned to teach each course. A similar practice commonly is used in determining topics for inclusion on examinations whether prepared for local or national use. These practices place sole responsibility for content decisions with those individuals who have advanced training and extensive experience in the area under consideration. While there are many advantages to this arrangement, there also are potential problems. One fear is that faculty will prepare curriculum materials or test items which are too difficult for individuals with much less experience in the specialty. A second is that topics covered may not correspond with the needs of students preparing for a career outside of that particular area. These concerns are both important issues in undergraduate medical curricula where specialists are responsible for teaching students whose needs are different from theirs both in terms of the amount to be learned and the use to which the knowledge will be put.

Many individuals have recognized this problem. Some (Page, et.al., 1976; Spivey, 1971) have analyzed the problem systematically and/or provided suggestions for checking tendencies toward teaching or testing material incompatible with student needs. Most solutions proposed depend on involving individuals with varying perspectives in the decision-making process. Spivey's approach to developing an undergraduate opthalmology curriculum is particularly noteworthy.

In preparation for developing end-of-year comprehensive examinations for evaluating medical students' performance at Southern Illinois University School of Medicine (SIU-SM), students, faculty and residents were asked to rate each basic science behavioral objective in terms of the importance for students to retain each competency for later use. This paper, a report to that project, has two major purposes. The first purpose is to describe the method itself

[1] This research was supported in part by Grant No. G007502225 from the Fund for Improvement of Postsecondary Education (DHEW). The authors would like to thank Kay Bleach for assistance in data collection and John Schmidt for assistance in data analysis.

Reprints can be obtained by writing to the first author at the following address:

Office of Curriculum Affairs and Educational Resources
Southern Illinois University School of Medicine
P.O. Box 3926
Springfield, Illinois 62708

as it provides a practical means for reviewing curriculum and test content and for creating test blueprints. The second purpose is to report comparisons of ratings by basic medical science faculty, clinical faculty and students; a reflection of the level of agreement about important basic science content. The specific research questions are spelled out in the results section of the paper.

METHOD

Rater opinions were expressed through use of a five point rating scale. The method section of this paper will describe the rating scale, characteristics of raters, and procedures for collecting the data.

Objectives Rating Scale

A description of the rating scale will be provided first followed by a discussion of the logic for using this scale. A five point rating scale was used. The descriptors anchoring each point on the scale were as follows:

A---Retention of this competency in its originally learned form and detail is essential for the duration of Sequence I.

B---Retention of this competency in its originally learned form and detail is highly desirable but only retention of the main idea is essential for the duration of Sequence I.

C---Retention of the main idea is desirable but not essential.

D---Important only to recall existence and to be able to look up when needed.

E---This objective is inappropriate for undergraduate medical education.

Raters were instructed to leave the item blank only if they had "absolutely no understanding of the terminology or the concept" and, therefore, had "no basis for judgement."

The focus of the rating scale was on what students should retain long after instruction was completed rather than on what they should be exposed to during instruction or be able to demonstrate immediately after completing instruction. The philosophy was that this approach would result in identification of the most important concepts and procedures to be maintained. These could then serve as the basis for comprehensive examinations, for curriculum evaluation, and for planning of rehearsal activities in later sequences to maintain needed competencies.

In designing the rating scale it became clear that the decision was not a simple yes-no decision for each curricular objective. Rather, the critical decision was to determine the form in which competencies should be retained. In some cases a competency must be retained in its originally learned form and detail. In other cases, retention of the main idea is enough. In still other cases it is only important that a student knows that a body of knowledge exists and knows where to look it up when needed.

Raters

Individuals representing four distinct perspectives were asked to rate objectives for each learning module. 1) Basic science faculty (N=84) had the greatest understanding and knowledge of the specific content of the basic science objectives. The module author and one other basic scientist familiar with the content were asked to rate each objective. 2) Physicians (N=72) who teach portions of the curriculum subsequent to the basic science portion comprised the second group. Their perspectives include recognition of competencies needed for successful completion of clinical instruction and knowledge of important competencies for effective medical practice. 3) Advanced standing students (N=94) were in an ideal position to relate the demands of their current instruction to competencies taught in the basic science portion of the curriculum. Second and third year students were asked to participate. 4) Residents in primary care specialties (N=62) were asked to participate. They represented the perspective of individuals who have recently completed undergraduate medical training and are currently in practice.

Procedure

A questionnaire package was assembled to be sent to all raters. The package contained a cover letter explaining the project and uses of the data, the rating scale instructions, a series of module objectives to be rated, optical scanning data sheets on which the ratings would be recorded and a pre-addressed stamped envelope.

Basic science faculty rated modules which they had authored and others in their area of expertise. Clinical faculty, students and residents rated modules which were randomly assigned to them. Basic science faculty were asked to rate objectives from an average of fifteen modules while the other raters were assigned to rate objectives from eight modules. Each module has an average of eight objectives, but there was wide variation in the actual number. Every module was assigned to be rated by two members of each group. Thus, ideally every objective would have been rated by eight different people.

Returned ratings were automatically key punched from the optical scanning sheets and a computer generated summary of ratings was obtained. The summary included rater information (rater ID, group affiliation, specialization, etc.), the name and number of the module being rated along with the individual ratings of objectives that were submitted. The computer summary, also, provided an average rating for each objective. This was obtained by assigning values of 4, 3, 2, 1, 0 to each rating category respectively with a 4 indicating that the competency should be retained in its originally learned form and detail. Blanks were disregarded in determining the average rating.

RESULTS

Return Rate

Questionnaire materials were sent to all raters in February. Three weeks later 58% had completed and returned their materials without benefit of a reminder. After two months had elapsed and one reminder had been sent, 59% had returned their materials. However, this figure is depressed dramatically by a 16% return rate of residents and a 39% return rate of volunteer faculty. Without these two groups, the return rate was 80% (basic science faculty 88%, clinical faculty 70%, students 63%).

Results suggest that this is a practical strategy for use in validating objectives when used with regular faculty and students. The rate of return could probably be improved if the number of modules assigned to each individual was reduced. The task for our raters required between one and three hours to complete.

Quality of the Rating Scale and Process

Inspection of the ratings printout suggested that the rating process was successful. All rating categories were used, indicating that each category was realistic (had applications). Further, the ratings results for objectives within modules differ widely. In 32.3% of the cases ratings for objectives within a module had a range of one category (e.g., A to B, C to D). In 31.8% of the cases the range was two categories. In 19.7% of the cases the range was three categories and the remaining 4% of cases had a range of four categories.

Viewed another way, in only 169 out of 1392 cases (12.1% of the cases) did the rater provide the same rating for every objective in a module suggesting that the rater might have been responding to the module topic rather than the objective or might have been uncritically assigning the same rating to all objectives for other reasons. This figure may exaggerate the number of cases approached in this way as some raters could have critically studied each individual objective in a module and determined that all deserved the same rating. Twenty-six of 168 raters (15.48%) provided the same ratings for all objectives within a module for 25% or more of the modules they rated. Five (2.98%) rated a majority of the assigned modules in this way. These findings suggest that the number of raters who either did not take the task seriously or responded to the module topic rather than the individual objective was rather small.

Proportion of Objectives Deemed Essential for Maintenance

The mean rating for each objective was used to determine what percentage of all curricular objectives were viewed as representing competencies essential to be maintained. For this purpose objectives with a mean rating of 3.3 or above were identified. Descriptively, a mean of three indicates that raters agreed it was essential for students to at least retain the main idea represented in the objective. One third (32.9%) of the objectives in the basic science curriculum were assigned an average rating of 3.3 or higher. The range was from 4.5% of Behavioral Science objectives to 68.2% of Introduction to Clinical Medicine objectives. These two were clearly extreme cases. All but two other curriculum segments had roughly 30 to 40 percent of objectives rated as representing competencies essential to be retained by this definition.

These findings suggest an important distinction between two types of teaching objectives; those designed simply to expose students to an idea, area, etc., and those designed to require students to learn and maintain some competency for a long period of time. The distinction seems obvious once given the data, but experience suggests that this distinction is not manifested in differential forms of teaching and testing for the two types of objectives.

Agreement Among Raters

Since ratings for each objective were obtained from representatives of four different groups it is possible to determine level of agreement among these groups, information which may be of use in future curriculum planning or test development efforts. In this paper we will address four questions: 1) Do module authors rate their objectives differently than do other raters? 2) Do basic science faculty rate basic science objectives differently than do clinical faculty? 3) Do students rate objectives differently than do faculty? and 4) Do individuals from the same specialty demonstrate a higher level of agreement than do individuals from different specialties? Inferential statistics have been used for these comparisons and statistically significant differences will be reported. However, since the number of observations sampled is quite large (900 to 13,500), differences which are of low magnitude and thus little utility may be statistically different. Our interpretations will take this into account and readers are informed so they may govern their own interpretations accordingly.

Authors were asked to rate 3018 of their objectives. Other raters provided 13,557 ratings of these objectives. The mean rating of these objectives by authors was 2.85, while the others provided a mean rating of 2.87. A comparison between group ratings was performed using the procedure described by Scheffe' (1959). This difference was not significant ($F' = .623$, $p \leq .05$). In addition, product moment correlations were computed between author ratings and ratings of others. Where there were two authors (70 cases) the correlation between the authors ratings was $r = .54$, $p \leq .001$. The correlation between an authors ratings and the ratings of other basic science faculty was $r = .19$ ($n = 942$, $p \leq .001$). The correlation of author's ratings with clinical faculty ratings was about the same ($r = .21$, $n = 2541$, $p \leq .001$). The correlation of authors and student ratings was $r = .27$ ($n = 2485$, $p \leq .001$). Thus, there was some agreement between authors and others on the importance of objectives; however, the correlations are not particularly high when compared with the intercorrelation between ratings of module coauthors. These results suggest that adding the ratings of others may indeed provide new information and perspectives.

The second question asked whether basic science faculty rate basic science objectives differently than do clinical faculty. The mean rating of objectives by basic science faculty ($n = 6198$) was 2.86. Removing the authors from this group does not change the mean. The mean rating of these same objectives by clinical faculty was 2.98. Using the Scheffe' method of contrasts, the differences are statistically significant ($F' = 43.88$, $p \leq .05$) but not large enough to be practically significant. The correlation between ratings of basic science and clinical faculty is .21 ($n = 954$, $p \leq .001$).

A related question asked whether students rated objectives differently than did faculty. The mean rating of objectives by students was 2.78 ($n = 4780$ ratings) while the average faculty rating ($n = 11{,}055$ ratings) was 2.91. This difference is again statistically but not practically significant ($F' = 32.61$ $p \leq .05$). The correlation between student and basic science faculty ratings was .22 ($n = 959$, $p \leq .001$) and the relationship with clinical faculty ratings was .30 ($n = 1303$, $p \leq .001$). As a basis for comparison the correlation between two sets of student ratings was .37 ($n = 1589$, $p \leq .001$). While these correlations are statistically significant they are small enough to suggest that the addition of student ratings does provide a different perspective.

The final question involved determining whether individuals from the same specialty demonstrate a higher level of agreement than do individuals from different specialties. The correlation between ratings by raters from the same specialty was .13 ($n=923$, $p \leq .001$). The correlation between ratings by raters from different specialties was .17 ($n=939$, $p \leq .001$). A test for differences between independent correlations indicated the correlations are not reliably different ($Z=1.01$ N.S.). This result suggests that, when it comes to rating undergraduate medical curriculum objectives, raters of the same specialty are as likely to disagree with each other as they are to disagree with raters from another specialty.

DISCUSSION

The procedure described appears to be a viable way for systematically evaluating curriculum and test content when precision outweighs the added costs. On the positive side, 80% of the full-time faculty and students cooperated even though it took up to three hours of their time. Those who cooperated appear to have taken the task seriously. Most seem to have critically reviewed each objective and assigned an appropriate, independent rating. There is little evidence of overgeneralization from module topic and/or surrounding objectives. There, also, is little evidence of random responding. All points on the scale were used suggesting that the scale points are realistic.

On the negative side this process is quite demanding of faculty, student and staff time when done on this scale. The time required for raters to complete the ratings has been discussed. In addition, much time is required to prepare the optical scanning sheets, to determine potential raters and to assign modules to raters. We suggest that future efforts be restricted to reviewing one to three curriculum segments at a time.

The findings suggest there are two distinct categories of curricular objectives which imply functional differences in teaching and testing. Some objectives serve solely to expose students to ideas and areas for future investigation. Other objectives (maintenance objectives) represent competencies that must be thoroughly learned and continuously practiced to ensure continuing competence. Review of current educational practice suggests that the two types of competencies are functionally treated as one. Teaching and testing appear to be geared to treating all objectives as maintenance objectives.

Finally, the results of this study suggest a moderate level of agreement among module authors, basic science faculty, clinical faculty and students regarding the importance of maintaining competencies represented in each objective. This finding suggests that involving individuals other than authors may provide useful added perspective for curriculum or test review and/or development.

REFERENCES

Page, E. B., Jarjoura, D., Konopka, C.D. Curriculum design through operations research. *American Educational Research Journal*, 1976, 13, 31-49.

Spivey, B. E. A technique to determine curriculum content. *Journal of Medical Education*, 1971, 46, 269-274.

Poster Session
1979

EVALUATION STRATEGIES FOR RESIDENCY TRAINING IN BEHAVORIAL PEDIATRICS, or, "OVERCOMING ORGANIZATIONAL HURDLES"

Pamela J. Trent, Ph.D.
Department of Psychiatry

Robert A. Hock, Ph.D.
Departments of Pediatrics and Psychiatry

W. Sam Yancy, M.D.
Departments of Pediatrics and Psychiatry

PURPOSE

Pediatric Medicine has been aware for some time of the central role it plays in the primary care of the American public. In keeping with this concern, the American Academy of Pediatrics has expressed the ultimate goal of pediatric training as providing physicians with the ability to assist children to realize their full potential. This requires training for the resident in "well" child care and attention to behavior and development problems and psychosocial issues (1,2,3). Although this training is espoused by the profession, two very important questions remain: 1. "How can behavior, development and psychosocial training be optimally incorporated into pediatric residency programs?" 2. "What is the best way to evaluate the overall training effects of these efforts?"

These questions are the focus for this problem-solving poster session. First, we will present the history and current description of a multi-disciplinary, objective-based pediatric residency training program in Behavorial Pediatrics and Developmental Medicine which is based on principles of learning and curriculum development as discussed in the medical education literature (4,5). Second, because the organizational features of the program constrain optimal overall evaluation, we will invite collegial discussion about the strengths of various evaluation designs. This poster session therefore describes the program, discusses our evaluation attempts, and asks specific questions regarding ways in which other programs are contending with organizational obstacles to evaluative research in medical education. This is a brain-storming session where the problem issues of valid and reliable measurement, research design, quantitative and qualitative data analysis, and interpretation of results are discussed in terms of pediatric residency training.

PROGRAM DESCRIPTION

As early as 1957, the Departments of Pediatrics and Psychiatry at Duke University Medical Center collaborated to establish a Behavioral Pediatric Developmental Medicine training program for pediatric residents. Since then, federal (NIMH) and private (W.T. Grant Foundation) funds have supported program development as well as consultation with medical education specialists (Educational Resource Group -- Psychiatry). In keeping with the consultants' recommendations, current training is designed according to sound principles of curriculum development. That is, objectives have been defined and teaching methods determined for three levels of learning: (a) acquisition of theory, (b) development of skills based on theory, and (c) application of theory and

---

Reprint Requests: Pamela J. Trent, Box 3903, Duke University Medical Center, Durham, N.C. 27710.

skills in the health care setting.

The behavior and development training program is apportioned into 6 training modules, each of which addresses theory, skill, or application. These modules are developed and supervised by an interdisciplinary team in the Department of Pediatrics. The team, directed by a pediatrician, consists of pediatricians, child and liaison psychiatrists, clinical psychologists, social workers, recreational therapists, and an educational consultant. Table I briefly describes each module's focus, how it matches with level of training, and which team members are responsible for it. Examples of complete modules will be available as handouts at the poster session.

As part of a 3 year pediatric residency curricula, the behavior and development modules are scheduled on predesignated rotations (e.g., surgery, outpatient clinic, intensive care nursery). (See figure 1 for sample schedule.) Except for the Behavior and Development Seminar module, training time for the approximately 30 residents must be completely individualized with one or two residents going through a module at a time. Although the program is required, there is no systematic regulation for attendance and patients' acute medical problems take priorty over learning activities. This places certain constraints on program evaluation -- particularly on how the overall training effect is evaluated.

EVALUATION

Individual evaluation of learning is built into each module with results used as feedback to the instructor and trainee. The desire, however, is to evaluate the overall effectiveness of the program with specific attention to:

1. Trainee perception of competence and relevance of training
2. Change in attitudes toward psychosocial aspects of patient care
3. Acquisition of skill and knowledge
4. Actual time spent in behavior and development training
5. Longitudinal effects of training.

This data would provide significant information for improving the training and for sharing with other programs.

An evaluative research design which approximates those required for generalizable quasi-experimental research (6) would be ideal but seemingly impossible in thise case where organizational constraints complicate evaluation design. Although data is collected with valid instruments, the lack of controlled research design inhibits causal statments about the relationship between overall results and program characteristics. That is, the same results could have occurred in the previous program, which was based on a less structured curriculum model with no rigorous evaluation. Or, residents may have acquired the theory, skills, and attitudes as a matter of maturation, regardless of the training itself. Another ideal evaluation strategy would be to use anthropological techniques (e.g., participant observation) to collect data directly in the health care milieu. This would provide sensitive data which would be difficult to obtain with paper and pencil measures or interview. However, problems militate against this strategy as against the quasi-experimental design. The problems are presented next.

PROBLEM STATMENT

As suggested throughout this paper, a number of organizational features inhibit ideal evaluative research on overall program effectiveness. These are listed here:

1. There are limited resources for conducting evaluation. The grant that

supports program development does not earmark money specifically for evaluation. Thus, the educational psychologist (with 10% time to devote), clinical psychologist (with 5% time), and the pediatrician (with 5% time), must work together to develop and administer measures, analyze data, and report results without the aid of a support staff or funds.

2. The diffuse scheduling of the training and the nature of the clinical environment make it impossible to control the training variables and to determine exact resident involvement.
3. Much of the training in the modules is organized around actual patient cases. This affects uniformity of training because the cases change from week to week. Although the objectives remain the same, the psychosocial characteristics of the cases obviously affect the learning associated with the module.
4. It is impossible to have a control group either from the Duke program or from another residency program in the vicinity. In terms of the Duke program, all residents are affected to a degree by the behavior and development training. It is impossible to determine the extent of the effect. In terms of other programs, there are differences in philosophy and objectives among schools. Hence, comparisons would only be artificial in the long run.
5. This is a three-year program, requiring that evaluation data be collected on a longitudinal basis. Too many data collections (e.g., 3 per year) would pose a hardship to the residents who are already overburdened with responsibilities and would discourage thoughtful responses. But too few collections (e.g., 1 a year) would reduce the strength of interpretation by accentuating the intervening variable of time.

These problems continue to impinge on evaluation efforts. For this reason, we invite discussion on ways to use the problems to our advantage or to avoid them in setting up a rigorous evaluation design. The specific questions we would ask are listed in Table 2.

REFERENCES

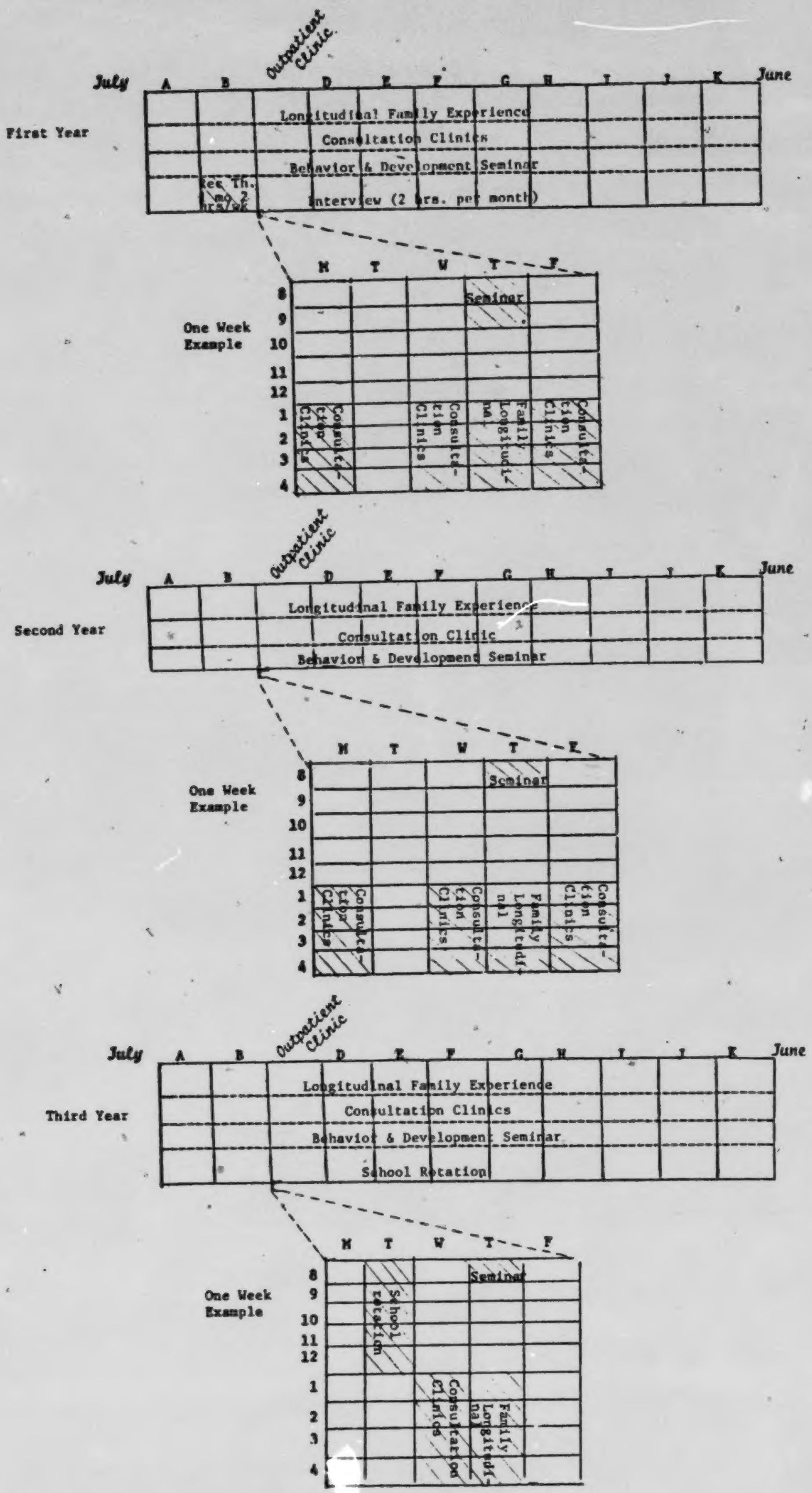
1. Deisber, R.W., Derby, A.J., and Stauman, M.J. Changing Trends in Pediatric Practice. Pediatrics, 25:711, 1960.
2. Hammond, K. How Will the Pediatrician Use Time Freed for Him? Clinical Pediatrics, 9:314, 1970.
3. Haggerty, R.J., The Changing Role of the Pediatrician in Child Health Care. American Journal of Diabled Children, 127: 545, 1974.
4. Guilbert, J.J. Educational Handbook for Health Personnel. Geneva, World Health Organization, 1977.
5. McGaghie, W.C., Miller, J.E., Sajid, A.W., and Telder, T.V. Competency Based Curriculum Development in Medical Education: An Introduction. Public Health Paper #68. Geneva, World Health Organization, 1978.
6. Campbell, D.P., and Stanley, J.C. Experimental and Quasi-experimental Designs for Research. Chicago: Rand McNalley, 1963.

Table 1

Training Modules for Behavioral Pediatrics and Development Medicine

Level of Learning	Module
THEORY	Behavior and Development Seminars Developed & conducted by pediatrician, child psychiatrist Focus: -normal & abnormal development -common psychiatric childhood problems -theoretical foundations of psychotherapies -family dynamics -physician-patient interaction
SKILLS	Training Program for Interview Process Developed & supervised by social worker, child psychiatrist Focus: psychosocial aspects of interviews
	School Rotation Developed & supervised by clinical psychologist, pediatrician, child psychiatrist Focus: -observation of healthy children in structured & unstructured situations -role of physician in schools -education & psychological testing -care of child with problemmatic behaviors
	Recreational Therapy Developed & supervisied by recreational therapists, child psychiatrist Focus: -emotional problems associated with illness -age appropriate coping behavior -observing hospitalized children in structured & unstructured play
APPLICATION	Consultation Clinics: (1) birth - 5 yrs., (2) 6 - 11 yrs., (3) 12 - 18 yrs. Developed & supervised by pediatrician, child psychiatrist, clinical psychologist Focus: -"well" child care -counselling parents on normal & abnormal devel. -care of child & family with behavior/development problems
	Family Longitudinal Experience Developed & supervised by pediatrician, child psychiatrist Focus: each trainee assigned 3 families to follow for 3 yrs. with specific supervision regarding behavioral & devel. issues.

FIGURE I
EXAMPLE OF A RESIDENT'S TRAINING SCHEDULE *


July
A
B
Outpatient Clinic
D
E
F
G
H
I
J
K
June
First Year
Longitudinal Family Experience
Consultation Clinics
Behavior & Development Seminar
Rec Th. 1 mo 2 hrs/wk
Interview (2 hrs. per month)
M
T
W
T
F
One Week Example
8
9
10
11
12
1
2
3
4
Seminar
Consultation Clinics
Family Longitudinal
Second Year
Consultation Clinic
School Rotation
School rotation
Third Year

* Shaded areas = Behavior & Development Training

Table 2

List of Specific Questions

1. Considering the large number of variables in a clinical training program, is it possible to do cause-effect evaluative research? What strategies could be used to enhance statements regarding causal relationship among training variables.

2. What are realistic goals for evaluating residency training in general? Is quasi-experimental research design possible for such programs?

3. What are the advantages and disadvantages of quantitative-experimental and qualitative-descriptive analyses for assessing training outcomes in a program with organizational features as described in this paper?

4. What are optimal measures for assessing attitudes toward psychosocial aspects of patient care? Is a disquised measure (e.g., case vignette) preferable to semantic differential, Likert scale, observation ratings, and open-ended questionnaires?

5. Does the number of questionnaires affect responses to program evaluation? For example, how do fatigue and test-retest factors affect the data?

6. How much importance do physicians assign to student evaluation of their instruction?

USING CONFIDENCE TESTING IN A PHYSICAL DIAGNOSIS COURSE

Robert M. Rippey, Ph.D., Department of Research in Health Education
Anthony E. Voytovich, M.D., Department of Medicine
University of Connecticut School of Medicine, Farmington, CT 06032*

Purpose

The transition of the medical student from the certainty of the basic science oriented world and National Board examinations to the uncertainty of clinical decision making is not an easy one. It is often formidable and at times traumatic. How can the student be better prepared for this change?

There may be a partial answer in confidence testing. Students are allowed to assign probabilities to the answers to their multiple choice tests instead of selecting only one option as correct. The student makes use of partial knowledge and is encouraged to say "I'm not completely certain."

Students who claim that they are certain when expert opinion is not so confident might be prone to make errors of clinical judgment known as premature closures (Voytovich, 1976). Thus, categorizing the student who "knows not and knows that he knows not" from the student who "knows not and knows not that he knows not" has clinical relevance. The purpose of this research is to demonstrate that the categorization can be done with reliability and validity in a way which students both accept and learn from.

Background

Unfortunately, implementation is not easy. Only a very limited number of scoring protocols are permissible. The problems of properly administering and scoring confidence tests have been detailed by Rippey and Smith (1979) and only recently have practical means for solving the technical problems been available (Landa, 1976; Rippey and Donato, 1978).

Early investigators claimed on theoretical grounds that confidence scored tests should be more reliable and valid than conventional multiple choice tests (Shuford, Albert, and Massengill, 1966). However, a review of early investigations attempting to confirm the added benefit of confidence testing showed that the yield was low and, if there were gains, they were seldom worth the additional effort to the students and the increased complexity of scoring (Echternacht, 1972).

Echternacht's report might have sounded the death knell for future research; however, help came from an unexpected source. The CIA commissioned the Rand Corporation to study the applications of confidence testing to the problem of determining methods of weighting intelligence estimates for accuracy. Rand demonstrated that this approach was not feasible (although certainly such a capability would have been useful in Iran of late). However the research did produce a method of adjusting scores for numerical estimates of realism (Brown and Shuford, 1973).

---

*Requests for reprints should be addressed to Dr. Rippey.

Prior to Brown and Shuford's research for Rand, it was possible to determine qualitatively whether or not a person was overvaluing or undervaluing information. Their research created a quantitative estimate of the actual amount of overvaluing or undervaluing. Thus, new scores of knowledge could be obtained which were adjusted for the realism (or guessing) component. This technique was employed in the present study and is described below.

Methodology and Instrumentation

Confidence testing depends in theory on a set of interlocking conditions which must be met at test time: 1) students are given the option of responding with probabilities, 2) students are told how they will be scored, and 3) reproducing scoring functions are used. The reproducing function is the key to the method.

Reproducing scoring functions maximize student scores _if and only if_ the probabilities which they assign to options reflect their true state of knowledge. For purposes of confidence testing, true state of knowledge has an operational definition. True state of knowledge is a conditional probability represented by the proportion of the time a particular probability is assigned to the correct answer. That is, if a person assigns a probability of say, .8 to numerous options on a variety of test items and is right in those instances .9 of the time, then true state of knowledge is said to be .9 given an assigned probability of .8. Moreover, the student has slightly undervalued. The ideal probability assessor is always right when a 1.0 is used, never right when a 0.0 is used and right half the time when a .5 is used.

Most students are not precise in estimating such probabilities. However, on the basis of a sequence of responses, a regression line can be computed which relates assigned probabilities to true state of knowledge or proportion of times right. If the slope of the regression line is less than one, the subject has overvalued information. If it is greater than one, the subject has undervalued.

Shuford, Albert, and Massengill (1966) proved that a reproducing function (in this case, $50 \log 3 P_k$ and where P_k is the probability assigned to the correct answer except where $P_k = 0$) will maximize the score for the test when $\bar{r} = \bar{p}$ for each item. The condition $\bar{r} = \bar{p}$ means that the student is assigning realistic probability estimates as answers. The condition $\bar{r} = \bar{p}$ implies that "guessing" is minimized. When a test is rescored using the regression estimates of P_k in place of the originally assigned probabilities, the score increases. The increase in score is due to the removal of errors of realism in assessing knowledge. This increase will be referred to as "realism improvement" or RI. Using the preceding function, the maximum score possible for a ten-item test would be 230 points where $P_k = 1.0$ for each item. Overall improvement in score (OI) is simply the initial number of points missed. The score improvement possible based on knowledge alone with realism errors removed then equals OI minus RI. This score is called knowlege improvement (KI).

The details of this procedure will be displayed at the proposed poster session. Samples of answer sheets, student instructions, samples of student responses and analyses of these responses, and graphs of payoff and realism functions will be displayed. Access to the adjunct computer programs will be made available. Persons interested will be able to take confidence tests, have them scored, and have interpretations returned to them.

The preceding description suggests that students may have to devote extra time and energy in responding to test items in this fashion, and that they run the risk of double jeopardy. Not only might they fail a test because of lack of knowledge, they run a second risk of being either too sure of themselves or too insecure. Considering these problems, student cooperation might be poor.

The objections were surprisingly easy to overcome by using a trick learned at an early age at the Elkhart County Fair. You offer two bingo cards for the price of one. If everyone receives two bingo cards, the probability of winning remains the same while the illusion of advantage is enhanced. Students were offered confidence testing as an option. They could respond conventionally, split their confidences, or both. Papers would be scored using both the conventional "number correct"and confidence KI score. Z-scores would be computed for the class and anyone better than 1.2 standard deviations below the mean by either method would pass the test. In reality this offer improved the changes of "winning" slightly. Needless to say, every student accepted a second bingo card.

The Sample

Seventy-six second year students at the University of Connecticut Health Center School of Medicine enrolled in Introduction to Clinical Medicine were administered three confidence scored tests during the course of the year. A very short test was given early to acquaint students with the system and to allow them to ask questions. The second test was a 49-item midterm. The third was a 97-item final examination. In addition to these tests, students also received ratings on their written case histories. Tests were scored and even-odd correlations computed for 1) conventional scoring, 2) raw confidence scores, 3) confidence scores adjusted for realism, 4) realism scores, and 5) angle of the realism function.

The final examination was further broken down into two approximately equal parts so that correlations between realism function angles of two halves could be obtained as an estimate of reliability. Next, a comparison of the five scores above was made between students ranked high and low on their case histories. Comparisons were made between the practice test, the midterm test, and the final test to see what types of changes occurred as a result of learning during the year's intervening experience.

Results

Means and standard deviations and reliabilities are shown in Table 1.

TABLE 1: SCORES AND RELIABILITIES FOR THREE TESTS

PRACTICE TEST N=68, 25 items Max=525	CK[1]	Confidence Score of Knowledge OP[2]	 AP[3]	RI[4]	Slope of Realism Function	AN[5] (ideal is 45°)
Reliability	not done	.31	.45	.48	Mean = 0.690 S.D. = 0.200	35.0°
MIDTERM N=78, 49 items Max=1127						
Mean	42.600	690.40	730.70	40.56	Mean = 1.099	47.7°
Percent	87.000	61.00	65.00	4.00	S.D. = 0.161	
S.D.	3.331	150.60	156.10	45.42		
Reliability	0.600	0.68	0.68	0.72		
FINAL N=76, 97 items Max=2231						
Mean	84.300	1345.80	1440.00	91.87	Mean = 1.089	47.4°
Percent	87.000	60.00	65.00	4.00	S.D. = 0.199	
S.D.	5.800	347.00	304.00	93.17		
Reliability	0.610	0.83	0.82	0.86		

---

[1] CK = conventional score of knowledge

[2] OP = overall performance = (max - OI)

[3] AP = adjusted performance = (max - KI)

[4] RI = realism improvement score (small values are desireable)

[5] AN = realism function angle

In the final examination, the reliabilities of the confidence scores, both overall and adjusted for realism are larger than and significantly different from the conventional scores at the .01 level. The realism score also has a significantly higher reliability which means that persons who correctly valued their information did so consistently on both halves of the test. The realism score however, looks only at the amount of correction applied and ignores whether the realism amounted to overvaluing or undervaluing. Although small RI scores are desireable, it is even more desireable to know that persons who overvalue on one occasion also overvalue on another. To test this, the final examination was split into two halves and the angle of the realism regression line computed for each half. The resulting reliability was .89. Thus, it appears that the angle of the realism line is remarkably reliable as an estimate of the tendancy to overvalue or undervalue. Second, the realism angle has a low linear correlation with knowledge improvement (r = .02, not significantly different from zero). Thus, KI and realism scores are measuring two independent test entities. This finding was supported by an earlier study (Rippey and Smith, 1979).

The next analysis examined the concurrent validity of the confidence measures compared with conventional multiple choice testing. A criterion was established by categorizing student clinical workups of a single patient into superior, average, and below average. T-tests were then performed comparing the high and low workup groups on the various conventional and confidence scores. Posters will be presented which will show that any of the three confidence measures seemed to discriminate about twice as well as conventional scoring.

The final question of learning had a very positive answer. Most students were able to learn to evaluate their level of mastery quite rapidly after the single practice test and the preliminary instruction. The average realism improvement score on the practice test was over five points per item. On the midterm it was less than one point per item. On the final it was less than one point per item (maximum = 100 points per item). The number of students overvaluing, undervaluing, and correctly valuing on the three tests is shown in Table 2.

TABLE 2: NUMBER OF STUDENTS OVERVALUING, UNDERVALUING, AND CORRECTLY VALUING ON THREE TESTS

	Overvalued	Undervalued	Correctly Valued
Practice	49	0	19
Midterm	2	16	61
Final	8	13	55

Although the realism means are close to the ideal, the fact of greatest educational significance is that not ALL students were able to pick up this skill rapidly and some did not have it by the end of the year. Of the group of 18 selected on the basis of clinical workups, two of the 18 substantially exceeded the one point per item norm of realism error on the midterm and one exceeded the norm on both the midterm and the final. Further analysis of the consequences of gross realism errors on clinical judgment should be interesting.

Conclusions

A high proportion of well motivated medical students seem to be able to improve their realism in assessing their levels of knowledge within a short time provided they are given adequate feedback. Reliability and validity of tests of knowledge can be improved by the application of confidence testing procedures. These procedures are feasible and acceptable to medical students provided they are administered under appropriate conditions. In addition, the confidence scoring procedure helps students express and learn to control their uncertainty. Information about tendencies to overvalue or undervalue information provide a new and independent dimension to test scores.

Several directions for the future are suggested. The relationship between confidence measures and the tendency to make premature closures in clinical situations needs to be clearly established. The characteristics of students who are resistent to improvements in their realism scores need further study. Teaching techniques need to be developed for improving the realism scores of the students having difficulty.

References

Brown, T. and Shuford, E. "Quantifying Uncertainty into Numerical Probabilities for the Reporting of Intelligence," a report prepared for the Defense Advanced Research Projects Agency, Santa Monica: The Rand Corporation, 1973.

Echternacht, G. "The Use of Confidence Testing in Objective Tests," Review of Educational Research, 42:217-236, 1972.

Landa, S. "CAAPM: Computer-Aided Admissible Probability Measurement on PLATO IV," a report prepared for Defense Advanced Research Projects Agency, Santa Monica: The Rand Corporation, 1976.

Rippey, R. and Donato, J. "Interactive Confidence Test Scoring and Interpretation," Educational and Psychological Measurement, 38:153-157, 1978.

Rippey, R. and Smith S. "Improving the Reliability and Validity of Confidence Scored Tests by Adjusting for Realism," Evaluation and the Health Professions, 1:100-109, 1979.

Shuford, E., Albert A., and Massengill, N. "Admissible Probability Measurement," Psychometrika, 31:125-148, 1966.

Voytovich, A. "An Analysis of Student Performance as Reflected in the Non-Threatening Audit of the Structured Medical Record," Academic Decision Making--Issues and Evidence, Washington, DC: AAMC, pp. 37-42, 1976.

PRECIS

PATIENT SIMULATORS IN CLINICAL TEACHING

MODERATOR: Donald A. Bosshart, Ed.D.
Northeastern Ohio Universities

EVALUATION OF A SELF-INSTRUCTIONAL TRAINING PROGRAM FOR SIMULATED PATIENTS

Evaluation of a self-instructional training program for simulated patients showed significant gains in performance from pretest to posttest. Implications are for increased effectiveness of such training programs in health professions education, as well as potential for such training in other areas.

AN INSTRUCTIONAL PROGRAM USING PATIENT INSTRUCTORS AS TEACHERS AND EVALUATORS

Non-physician patient instructors can function in the multiple roles of patient, teacher and evaluator. They provide individualized student monitoring with immediate feedback, and objective reliable evaluation.

A COMPARISON OF INSTRUCTIONAL TECHNIQUES: INTERNAL-EXTERNAL VS EXTERNAL ONLY

Two instructional models for effective teaching of the genital-rectal examination to beginning medical students are compared. Dramatic results indicate that the teaching method in which a non-physician instructor receives the examination is more effective than that in which a physician instructor only observes the examination.

68

EVALUATION OF A SELF-INSTRUCTIONAL TRAINING PROGRAM FOR SIMULATED PATIENTS

Barbara F. Johnson, Ph.D., Linda Perkowski,M.S. and Roger A. Lanier, Ph.D.

University of Texas Medical Branch at Galveston

Background

For several years, simulated patients have been used at the University of Texas Medical Branch (UTMB) in Galveston for teaching students medical interviewing skills. Due to the increasing use of simulated patients, their training was taking staff members an increasing amount of time while the quality of the training was varying significantly with availability of time and the individual simulated patient's past experiences. In order to improve the quality and consistency of the training of simulated patients while reducing the amount of staff time spent, the Division of Biomedical Communications designed, developed, and produced a self-instructional, mediated training package to provide a high level of standardized instruction for all simulated patients. Simultaneously, the provision of this training through other media allowed staff members more time to deal with higher levels of training of simulated patients, as well as added time for the complex recruiting, scheduling, and evaluating of simulated patients.

The purpose of this study was to evaluate a training program for simulated patients on two parameters; first its effectiveness in providing basic communication and evaluation information and skills to the simulated patient, and second its effectiveness in reducing staff time spent in these activities.

Review of Literature

A simulated patient is an individual who has been trained to accurately replicate a patient or any aspect of a patient's illness, depending on the educational need (1). In the past decade, the use of such individuals hired or volunteering to play the roles of patients--simulated patients--has grown steadily in medical education (2). These simulated patients give the student the opportunity to practice interviewing, and to get feedback on his/her communication and interpersonal skills. The advantage in using simulated patients lies in the standardization of the experience for the novice interviewer, a rationale which is supported by research studies (3-7).

Hobard (8) states that standardization of simulation is essential to its usefulness as an evaluation tool. Other studies (3,4,5,9,10) show that training of the simulated patient is essential to achieve the standardization.

Methodology

At UTMB, approximately 1,000 hours of interviews involving simulated patients are conducted each academic year, involving a cadre of 60 simulated patients. These individuals are recruited from the campus or community and, in

---

1. Requests for reprints may be sent to Barbara F. Johnson, Director, Division of Biomedical Communications, University of Texas Medical Branch, Galveston, Texas 77550

general, must be taught interview and evaluation skills. To date, this instruction has been conducted as a part of the overall function of the Simulation Coordinators, a staff of two individuals involved fulltime in recruiting, training, scheduling, conducting and evaluating interviews for varied student populations in the Medical Branch. These two staff members spent approximately 12 months developing a modular training program which would provide systematic training in simulation and evaluation of interviewing skills.

The materials developed and administered in the November 1978 to May 1979 training period consist of 10 units, containing 8 training videotapes and 10 print modules. Videotapes cover the following facets of simulations, interviewing, and evaluation:

- overview of the simulated patient program at UTMB;
- demonstration of appropriate uses of the simulated interview facility;
- group discussion on simulation by experienced simulated patients;
- group discussion of evaluation by an experienced simulated patient and faculty members;
- interviews by strdents with simulated patients, including the verbal evaluation followup sessions.

The print units that reinforce and supplement the taped instruction add the logistic detail of the simulated patient program, simulation theory and essence of the literature, commentary on taped interviews and evaluative sessions, and basic instructions on evaluation (both written and oral, since both are used in the simulated patient program). Printed modules contain objectives, content, and self-assessment tests. Trainees retained all print materials.

Staff members also developed a practice interview for inexperienced trainees in which a staff member plays the interviewer and provides immediate feedback on the trainee's performance.

Sample. At the beginning of the 1978-79 year, the pool of simulated patients included 41 who had had experience in prior years, and 18 who had applied for inclusion in the simulated patient program in the new academic year, but had had no prior experience. Both experienced and inexperienced individuals went through the training program, with a total of 41 completing the training. All simulated patients at UTMB are paid a minimal fee for their time as simulated patients, and the group going through the training program was paid a total of $20 per person on completion of the posttest, a figure predetermined on the "average" length of time (6 hours) an individual would devote to the training sessions.

Research design. All trainees, both experienced and inexperienced, were randomly assigned to either the experimental (N = 22) or control (N = 19) group for training.

All trainees completed the pretest. The experimental group then proceeded with the training materials on an individualized basis, completing a self-determined number of units in any one training session. At the end of his/her training program, each trainee was given a posttest and an opinionaire. Training began in November, 1978, and was conducted in the same room for all trainees. In February and March, 1979, the control group trainees again took the posttest, and filled out the opinionnaire. All trainees had completed the training materials by May, 1979.

Instrumentation. The pre-posttest for the training program is a 41-item written test, including fill-in-the-blank, true-false, listing, labeling, and short constructed response questions. Items are objective based and weighted; total possible score is 100. All tests were graded by an independent evaluator. With the posttest, an opinion instrument is used. It contains 17 statements about the program. Responses were on a five-point scale, from "strongly agree" to "strongly disagree." A line for comments was included. Additionally, the simulated patient was asked to rate each section (i.e., each print topic and each videotape, by name) of the program as H (helpful) or N (not helpful). The opinionnaires were devised for information on which to base revisions of the materials.

Results

In order to determine whether exposing simulated patients to a self-instructional training program significantly changed the amount of knowledge the simulated patient obtained and retained, t tests for independent observations were performed. A large significant difference ($t = 4.35$, df = 39, $p < .001$) was found between the posttest scores of the Trained (Experimental) ($\bar{x} = 83.00\%$) and Untrained (Control) Simulated Patients ($\bar{x} = 67.35$). When the group pretest scores were examined, no significant difference was observed ($t = 0.37$, df = 39, $p > .05$) between the Trained ($\bar{x} = 63.59\%$) and Untrained Simulated Patients ($\bar{x} = 61.63\%$). Thus, as would be expected, the gain score from pretest to post-test for the Trained Simulated Patients ($\bar{x} = 20.00\%$) was significantly greater ($t = 3.25$, df = 39, $p < .01$) than the Untrained Simulated Patients' score ($\bar{x} = 5.05\%$).

Since it was impossible to control for the number of hours of experience as a simulated patient the individuals received after being randomly assigned to groups due to the need to schedule simulated patient interviews during the training period, t tests were performed on this variable for the pretest, posttest, and the amount of experience gained during training. The differences in hours of experience were nonsignificant at the administration of the pretest ($t = 1.49$, df = 39, $p > .05$), the posttest ($t = 1.69$, df = 39, $p > .05$),and for the amount of experience gained during training ($t = 1.50$, df = 39, $p > .05$). The trained simulated patients tended to have larger average amounts of experience at the pretest ($\bar{x} = 22.04$) and at the posttest ($\bar{x} = 26.72$) than the untrained simulated patients ($\bar{x} = 9.73$ and $\bar{x} = 11.78$, respectively), but this was primarily the result of one individual who was randomly assigned to the training group with over 100 hours of experience. The trained simulated patients ($\bar{x} = 4.68$) also tended to gain more experience from the pretest to the posttest period than the untrained group ($\bar{x} = 2.05$). These results are summarized in Table 1.

Conclusions

The Trained and Untrained Simulated Patient Groups failed to show significant difference on any of the comparisons, except the absolute difference of their posttest scores, and their gain in performance from pretest to posttest. This would suggest that the training program for simulated patients significantly affected the amount of knowledge which was obtained and retained as a result of the group receiving the training. This finding supports the effec-

TABLE 1

A COMPARISON OF THE KNOWLEDGE GAINS OF TRAINED AND UNTRAINED SIMULATED PATIENTS*

Comparison	Trained (N = 22)		Untrained (N = 19)		t test for Independent Observations
	Mean	Standard Deviation	Mean	Standard Deviation	
Pretest	63.59	19.07	61.63	13.11	0.37
Posttest	83.00	9.67	67.35	13.78	4.35***
Gain Score	20.00	18.99	5.05	6.70	3.25**
Hours of Experience at Pretest	22.04	33.53	9.73	13.67	1.49
Hours of Experience at Posttest	26.72	35.64	11.78	15.33	1.69
Gain in Hours of Experience	4.68	6.93	2.05	3.37	1.50

* *All t tests are two-tailed; degrees of freedom are 39.
** $p < .01$
*** $p < .001$

tiveness of the training program in providing basic communication and evaluation information to the simulated patient. The effectiveness of the program in improving skills in the simulated patient is currently undergoing evaluation. A further observation is that, in the preparation of a trained, inexperienced simulated patient for a simulation activity, staff time was reduced.

Implications

The results of the evaluation study have implications both within and outside the Medical Branch. Within the Medical Branch, among the most important implications are:

1. With an established and ongoing training program for simulated patients, faculty members are able to view training materials and obtain the "definition" of a simulated patient. This provides the faculty with clear knowledge of what the simulator is or is not trained to do, and to work within or beyond the program without confusion. Additionally, with the knowledge of the basis for the simulated patient's feedback, the faculty member can coordinate his/her own feedback to the student, modifying it as it seems appropriate.

2. It is possible, as greater numbers of new simulated patients are introduced, that the time saved by a packaged, self-administered training program will save sufficient time to cut personnel costs. The need for additional staff members managing in a simulated patient program may be delayed by the availability of self-instructional training. There may be actual budgetary savings, in addition to personnel resource, or time, savings.

3. Where training had previously been accomplished through the instruction by a staff member to each simulated patient at the time of his/her simulation, the program now is available to new simulators at the best time suited to them, without interruption of other simulation tasks. Coordinators had dealt concurrently with all simulation functions, such as recruiting, scheduling, supervising, and evaluating ongoing interviews, as well as with training. Implication of the new training program is that instructional activities will be more appropriate to the individual, smoother in pace, and will provide a closely standardized training and simulation experience.

4. With more time available during the supervision and evaluation of simulated patient interviews, the coordinators are able to concentrate on refining performance and evaluation skills of the simulated patients. This freedom to observe and evaluate is anticipated to enhance simulation skills as a higher level of training takes place.

5. The results of the evaluation of this training program in communication knowledge and skills suggest that it is possible to provide packaged, modular training programs for nonprofessionals to improve their basic knowledge of interpersonal skills. The positive results of this program indicate a need to study similar programs designed, developed, and produced for the improvement of interpersonal skills in other critical areas such as hospital staff, office staff, etc.

The effectiveness of this training program for simulated patients has implications for institutions other than the University of Texas Medical Branch. Although the Medical Branch has a unique and unusually large simulated patient program, the desirability of standardized and time-saving training for simulated patients extends to other programs also. The same advantages that are accruing to the Medical Branch are available to other institutions with the development and implementation of a training program designed to meet their individual needs.

References

1. Barrows, H. S. *Simulated Patients (Programmed Patients)*. Spingfield, Il.: Charles C. Thomas, 1971.

2. Collart, M. E. Human simulators as teachers: A guide to the application of an effective simulation strategy. *Educational Technology*, 1979, *19*, 7-14.

3. Helfer, R. E., Black, M. A., & Teitelbaum, H. A comparison of pediatric interviewing skills using real and simulated mothers. *Pediatrics*, 1975, *55*, 397-400.

4. Adler, K. Simulated patients for communication research. *Journal of Medical Education*, 1977, *52*, 151-153.

5. Froelich, R. E. A course in medical interviewing. *Journal of Medical Education*, 1969, *44*, 1165-1169.

6. Ward, N. G. & Stein, L. Reducing emotional distance: A new method to teach interviewing skills. *Journal of Medical Education*, 1975, *50*, 605-614.

7. Jason, H., Kagan, N., Werner, A., Elstein, A., & Thomas, J. B. New approaches to teaching basic interviewing skills to medical students. American *Journal of Psychiatry*, 1971, *127*, 1404-1407.

8. Hoban, J. D. Successful simulations for health education. *Audiovisual Instruction*, 1978, *23*, 20-22.

9. Werner, A., & Schneider, J. M. Teaching medical students interactional skills. *The New England Journal of Medicine*, 1974, *290*, 1232-1237.

10. Helfer, R. E., Black, M. A., & Helfer, M. E. Pediatric interviewing skills. *American Journal of Diseases of Children*, 1975, *129*, 1053-1057.

AN INSTRUCTIONAL PROGRAM USING PATIENT INSTRUCTORS AS TEACHERS AND EVALUATORS

Paula Stillman, M.D., Jane Ruggill, Paul Rutala, M.D., Darrell Sabers, Ph.D.
University of Arizona College of Medicine, Tucson, Arizona.

The ability of a physician to acquire accurate and complete data from a patient is crucial to making clinical decisions. The past decade has witnessed the development of innovative techniques to evaluate patient assessment skills. Non-physician instructors have often assumed central roles in these evaluation models. Helfer and his colleagues developed a teaching program which utilized trained mothers as instructors and evaluators of pediatric interviewing skills.[1] In an adaptation of Helfer's work, Stillman and her colleagues expanded the use of the trained mothers to evaluate the process as well as the content of the medical interview.[2] Parallel advances were made in the teaching and evaluation of physical examination skills. By eliminating the variable of patient behavior, Barrows succeeded in providing a well-controlled objective test of the student's clinical performance which served as an adjunct to traditional neurological instruction and evaluation.[3] Kretzschmar developed an instructional system which employed Gynecology Teaching Associates (GTAs) who functioned as both patients and instructors for repeated pelvic examinations performed by medical students.[4] Frazer and Miller reported preliminary work in which non-physician practical instructors were utilized to teach first-year medical students basic physical examination techniques.[5] Anderson and Meyer developed a program which utilized chronically-ill patients to teach medical students the technical aspects of the neurological, respiratory and musculoskeletal examinations, providing students the opportunity to practice examination skills and identify positive findings in a low-anxiety setting.[6]

Stillman and her colleagues expanded the role of the asymptomatic non-physician patient instructors (PIs) to include that of evaluator of the normal physical and neurological examinations.[7] At the completion of an examination, the PI assessed an examiner's performance using objective evaluation instruments and provided immediate corrective feedback to those examiners who failed to meet pre-established criteria. In general, these assessment strategies served to evaluate a student's proficiency in examination technique rather than provide an objective evaluation of his ability to identify and describe a patient's abnormal findings.

In an effort to evaluate a student's diagnostic ability, an evaluation model utilizing non-physician patient instructors (PIs) with stable positive cardiac or pulmonary findings on physical examination has been developed by faculty at the University of Arizona College of Medicine. The PIs use their own bodies as teaching material and are trained by physicians to function in the multiple roles of patient, teacher and evaluator of physical diagnosis skills.

This work was supported by grants from the National Fund for Medical Education and the Department of Health, Education and Welfare.

Address reprint requests to: Paula L. Stillman, M.D., University of Arizona College of Medicine, Tucson, Arizona 85724.

Development of Performance and Content Checklists

Performance checklists which detail essential examination maneuvers have been developed for each specialty examination. Items included on the final performance checklist reflect the concensus of both a specialist and general internist. For scoring, equal weight has been assigned to each item on the performance checklist. Although the information gained from performing a specific maneuver may vary from patient to patient, the execution of each maneuver is important for thoroughness. Performance scores are reported as percent of the total items performed correctly.

A content checklist for recording the patient's findings on physical examination has been developed for each specialty examination. It includes all physical abnormalities of a given organ system which an examiner should be able to describe after performing the maneuvers on the performance checklist. For scoring purposes, certain items on the content checklist are weighted more heavily than others, depending on their importance to the individual PI's diagnosis. Content scores, expressed as a percent, compare the number of points obtained by the examiner to the total points available. If specific physical findings of a PI are variable, the PI is taught to change the keyed response on his content checklist as his findings change. For example, a PI with asthma auscultates his own lung field prior to and at the conclusion of each student's examination to determine the site(s) of audible wheezes.

Selection of PIs

The PIs are hired by the College of Medicine as salaried staff members. A PI need not be medically sophisticated but must be sufficiently intelligent to master a new vocabulary. If a PI's disease is chronic and disabling, he must have acknowledged the fact that he will never be healthy, yet must be willing to learn more about the course of his disease and its sequelae. Physical stamina is also a requirement. Because a PI's findings must be evident to a student, a patient completely controlled by medications is not an optimal candidate. To ensure accurate assessment of student performance, the PIs should have relatively stable findings.

Training of PIs

The initial training session is conducted by a specialist who demonstrates and reviews each item on the performance checklist. PIs with findings in the same organ system are trained simultaneously and are taught the techniques and subjective sensations of each examination maneuver and the significance of their individual findings. Subsequent sessions are conducted by a general internist who demonstrates variations in approach and examination style so that each PI becomes familiar with a range of acceptable technique. Initial teaching sessions are audiotaped for later review. A lay language instructor's manual has been written to serve as a home study reference for each PI. Additional reference material such as physical diagnosis, anatomy, physiology, and internal medicine textbooks are available as needed.

To establish reliability estimates for performance scores, several students examine each PI while the other PIs and the program coordinator observe. At the completion of each examination, the PI who serves as the patient and the observers score the examination checklist and record items performed by the student. Training of the PIs continues until interrater reliability assessed by a phi coefficient comparing items checked by each PI and the program coordinator is consistently high (average phi coefficient was .89 for the cardiovascular examination and .77 for the pulmonary examination). Higher interrater reliability is difficult to establish because some aspects of the physical examination are subjective and only the "patient" and not the observing PI or physician can tell whether the procedure is performed correctly. The difficulties encountered in establishing higher reliability present additional evidence of the shortcomings of traditional methods of assessment.

During the final training sessions, student volunteers examine each PI in the presence of the program coordinator and receive feedback from the PI on examination technique and diagnostic accuracy. Similar sessions are scheduled until each PI is judged capable by the program coordinator of providing quality teaching and instructive feedback to an examiner. Total training time is approximately 25 physician contact hours per pair of PIs in each specialty area. Once the PIs are trained they function independently of direct physician input. Periodic spot checking by the program coordinators is necessary to ensure the quality of the feedback.

The Program

Three cardiovascular and four pulmonary PIs were utilized to evaluate the clinical skills of three classes of medical students. Students in the classes of 1980 and 1981 received copies of the performance checklists prior to their first encounter with each PI. Students in the class of 1979 were not given performance criteria in advance. Content checklists were not distributed to any students prior to their evaluation sessions.

Each student was randomly assigned to examine a PI in each specialty area. Following the examination, the PI scored the performance checklist evaluating the thoroughness of the student's examination, and the student completed the content checklist identifying and describing the PI's positive findings. The PI then assumed the role of "teacher", reviewing the student's examination techniques via the performance checklist and diagnostic accuracy via the content checklist. Maneuvers which the student omitted or performed incorrectly were reviewed; physical findings which the student identified or described inaccurately were discussed. The PI provided feedback to the student on his interpersonal interaction within the examination setting and suggested methods to enlist a patient's cooperation and increase his comfort. The student repeated relevant portions of the examination guided by the PI. Mastery level passing criteria for performance and content were established for both specialty examinations. Students failing to meet these pre-set criteria were required to repeat the examination on another PI at a later date.

Results

Effect of prior knowledge of performance criteria

There was a difference (p=.01) in the mean performance scores achieved by students who received the cardiovascular performance checklist prior to their first cardiac examination (classes of 1980 and 1981) and those who did not (class of 1979). However, no significant differences were found in the mean cardiac content scores for the three classes. Prior knowledge of pulmonary performance criteria affected both pulmonary performance and content scores (p=.01) although the differences are much greater between performance scores than content scores.

77

Correlation between performance and content scores for each specialty examination - A correlation between performance and content scores for each specialty examination was computed separately for each class of students. Although five of the six correlations were positive, none was significant even at p=.10 level. The only significant correlation (p=.01) was negative.

Correlation between cardiovascular and pulmonary scores

A correlation between the scores on the cardiovascular and the pulmonary performance and content checklists was computed separately for each class. The only significant correlation was positive and was found between cardiovascular and pulmonary content scores for the Class of 1981.

Effect of the program on student learning

Each student who failed to achieve pre-set mastery criteria on his first cardiovascular or pulmonary examination was scheduled to perform a repeat examination on a second PI at a later date. A total of 50 students were required to repeat the cardiovascular examination and 73 students were required to repeat the pulmonary examination. The gains in both cardiovascular and pulmonary content scores were significant (p=.01). A recurring problem in the interpretation of gain scores is that students failing a first examination are expected to increase their scores on a second examination due to regression toward the mean. However, the gains in content scores were greater than could be attributed to statistical regression because the content means for the subgroup of students taking the second examination exceeded the means of all students on the first examinations. There were no significant gains in performance scores.

Student evaluation of the program

Approximately 80 percent of the students returned questionnaires evaluating the program. Student response to the PI program was overwhelmingly positive. Over 98 percent of the students recommended that the evaluation and feedback sessions with the pulmonary and cardiovascular PIs be continued for future classes of students. Ninety-five percent of the students rated the sessions as extremely helpful in identifying their strengths and weaknesses in examination technique and diagnostic accuracy. Students appreciated the relaxed atmosphere, the opportunity to examine cooperative patients with known findings, and the instructive and immediate feedback provided by the PIs. Suggestions for improvement of the program included a request that content checklists be distributed prior to a scheduled examination and that additional patients with diverse findings in each specialty area be trained.

Discussion

Students who had seen the performance checklist prior to their evaluations had significantly higher performance scores than those students who had not. This result was expected. Once reasonable passing performance criteria are clearly defined, competent motivated students are able to meet them. Prior knowledge of the performance criteria does not appear to have a consistent effect on content scores. For this study, none of the students had seen the content checklists in advance. To determine the effect of prior knowledge of content items, the content checklist will be distributed to future classes prior to their initial evaluations.

The data support the contention that a thorough examination (i.e., high performance score) does not imply that the examiner is able to identify and accurately describe physical examination findings. Students who performed the most thorough examinations were not necessarily those who could describe the patient's findings most accurately.

There appeared to be no consistent correlation between the cardiovascular and pulmonary performance scores or between the cardiovascular and pulmonary content scores obtained by students. This implies that there is little transfer of skills between the specialty examinations and each must be taught and evaluated separately.

Students who failed to achieve passing criteria on their initial evaluations appeared to learn during their teaching and feedback sessions with the PI. Significant gains were found in content scores but relatively high initial performance scores (mean=93%) may have limited further improvement.

What are the potential advantages and disadvantages of PIs as compared to more traditional methods of assessing an examiner's data collection skills and diagnostic accuracy? Short answer multiple-choice examinations, though reliable and relatively easy to design and score, are more useful for evaluating the recall and interpretation of facts than for predicting professional competence. Although patient management problems can be designed to challenge some aspects of clinical reasoning and problem-solving, their validity as monitors of clinical competence has never been established. Computer simulations, while measuring some aspects of problem-solving and clinical reasoning, eliminate the very important variable of the examiner's personal interaction with the patient.

The advantages of using asymptomatic persons who can simulate physical findings as part of the teaching and evaluation situation has been eloquently summarized by Barrows.(8) Symptomatic PIs appear to possess many of the advantages of asymptomatic simulated patients and occupy a position even closer to a real patient encounter because they indeed are patients. In addition, they may have true physical findings such as murmurs, funduscopic changes, and organomegaly which cannot be simulated. PIs can function as an integral part of a competency-based clinical evaluation. As part of a formative evaluation, PIs can provide each student with individual feedback about his performance so that areas of weakness can be identified and corrected. By analyzing the checklists of many students, common errors and misconceptions may be identified and the faculty can be provided with objective evidence of areas in which improvement in instruction is necessary. As part of a final evaluation, PIs can assess student performance at the completion of a course or training period and provide final certification of each student's skills. Students achieving passing criteria are permitted to progress to the next level of training.

There is no evidence that the current design of the PI program allows for the prediction of the performance of students in a real patient setting. Rather, the program is an attempt to quantify some of the critical components that are integral parts of clinical competence. The PI program is presently under development in other specialty areas including rheumatology, neurology and ophthalmology. In addition, techniques similar to those developed for the students are now being used to evaluate housestaff.

Conclusions

Symptomatic, non-physician patient-instructors (PIs) have been trained by physicians to function in the multiple roles of patient, teacher and evaluator. PIs provide immediate constructive feedback and objective reliable evaluation of students' physical diagnosis skills, and ensure that pre-set mastery competencies are achieved. Data indicate that student learning occurs during these teaching and evaluation sessions.

References

1 Helfer, R.E., Black, M.A., Helfer, M.E. Pediatric interviewing skills taught by non-physicians. Am. J. Dis. Child., 129:1053-1957, 1975.

2 Stillman, P.L., Sabers, D.L., Redfield, D.L. Use of trained mothers to teach interviewing skills to first-year medical students: a follow-up study. Pediatrics, 60:165-169, 1977.

3 Barrows, H.S. Simulated Patients. Springfield: Charles C. Thomas, 1971.

4 Kretzschmar, R.M. Evolution of the gynecology teaching associate: an education specialist. Am. J. Obstet. Gynecol., 131:367-373, 1978.

5 Frazer, N.D., Miller, R.H. Training practical instructors (programmed patients) to teach basic physical examination. J. Med. Educ., 52:149-151, 1977.

6 Anderson, K.K., Meyer, T.C. The use of instructor-patients to teach physical examination techniques. J. Med. Educ., 53:831-836, 1978.

7 Stillman, P.L., Ruggill, J.S., Sabers, D.L. The use of practical instructors to evaluate a complete physical examination. Eval. Heal. Prof., 1:49-54, 1978.

8 Barrows, H.S. Simulated patients in medical teaching. Can. Med. Assoc. J., 98:674-676, 1968.

A COMPARISON OF INSTRUCTIONAL TECHNIQUES: INTERNAL-EXTERNAL VS EXTERNAL ONLY

Walter Gerber, M.D. (University of Iowa Hospitals and Clinics)*;
Sandra Matthes, M.A.T. and Mark Albanese, M.A.
(College of Medicine, University of Iowa)

PURPOSE: Effective teaching of clinical examination techniques to beginning medical students differs greatly when internal as opposed to external structures are the object. The traditional "preceptor external observation" (PEO) method has four main components: the preceptor 1) demonstrates appropriate techniques on real patients or on medical students themselves; 2) observes the students as they imitate his actions; 3) judges the quality and accuracy of the examination on the basis of the students' statements as to what they claim is found; and 4) provides feedback to the students relative to the accuracy of their observations. The teacher is not always able to verify accurately the performance of the student by any direct visual or tactile criteria of his own. Despite this limitation, however, most students are taught pelvic and male genital-rectal examination by PEO methods.

An alternative approach was developed at The University of Iowa Hospitals and Clinics for teaching the pelvic examination.[1] This new teaching model has been adapted for the male genital-rectal teaching program. It includes internal as well as external observation by a non-physician teaching-associate simulated patient trained to recognize the various components of an examination being performed on himself.[2,3] The elements of this internal-external observation (IEO) model involve the following: 1) two teaching-associates, one serving as the instructor and one as the patient, demonstrate appropriate exam techniques to a group of four students; 2) the team verifies the accuracy of each student's examination based on what the "patient" internally experiences and on what both instructors externally observe; and 3) the team provides immediate advice to the student regarding his/her technique. IEO preparation requires the hiring of persons to serve as patients (as well as ongoing supervision from a staff urologist), whereas the PEO uses students as peer patients (or actual hospitalized patients). This paper reports the results of a prospective, randomized comparison of these two instructional models for the teaching of the male genital-rectal examination.

METHODS: The one hundred seventy-nine sophomore medical students enrolled in our Introduction to Clinical Medicine course were randomly assigned to one of two treatment groups for their genital-rectal training program. Ninety-two were trained by the internal-external observation (IEO) method in afternoon or evening sessions lasting approximately two hours. The details of the training of the teaching associates and their interactions with the students are described elsewhere.[2,3] The students were encouraged to practice until they felt comfortable with their ability to perform the exam.

Eighty-seven students were trained by members of the Department of Urology (both senior and resident staff) by means of the PEO approach in two consecutive evening sessions that also lasted approximately two hours. The urologists had been instructed in advance regarding minimum standards of information which they were expected to impart to their individual groups. These standards included all points to be covered by the teaching-associates and those to be evaluated later. A male student volunteer was sought from each group of four

*Reprints: Department of Urology, University of Iowa, Iowa City, Iowa 52242 to serve as the patient during the demonstration by the physician preceptor. Females in the group examined this male; the males in the group then examined each other. Females were not examined by their male peers. Students were encouraged, but not coerced, to participate; anyone, male or female, who felt uncomfortable had the option of withdrawing from the group at any time without penalty.

Instruction in the evening was utilized by the urologists for their PEO instruction because of their busy daytime clinical schedules. In an attempt to assess fatigue factors, IEO sessions were carried out both in the afternoon as well as in the evening.

The objective of both instructional models was to train sophomore medical students to perform the technical aspect of the genital-rectal examination at the highest level of accuracy and thoroughness possible and to provide them with interpersonal skills needed to establish comfortable communication with their patients. The fact that the students were participating in a research study was minimized at all times.

At the conclusion of the training, sixty-one students were selected at random to participate in an evaluation of their skills. The following groups were sampled (scheduling problems resulted in slightly unequal numbers): 49 students trained by IEO (23 from afternoon and 16 from evening sessions) and 22 trained by PEO during the evening. These students performed a genital-rectal examination on one of two teaching associate evaluators who were chosen for having the most experience in the IEO system. An evaluation form consisting of 27 items (see handout, Appendix I) was used to rate the thoroughness of the examination according to objectives previously established.[4]

To ensure that the evaluators would use as their criteria only the thoroughness of the examination they were blinded as to which groups the students belonged. No IEO-student was evaluated by the teaching-associate who had instructed him/her.

RESULTS: In Appendix II we have tabulated the interpersonal-skill ratings of the IEO and PEO groups, and in Appendix III the technical-skill ratings. The differences between the two instructional models were statistically significant in both areas in all skills assessed.

No differences were noted between IEO groups taught in the afternoon as opposed to those taught in the evening sessions. And no differences could be found between PEO students trained on different nights. A significant fact is that all of the students scheduled for IEO training participated; of the 87 students scheduled for PEO training, 18 (20.7%) either cancelled in advance or simply did not appear for their training due to personal reasons.

DISCUSSION: The IEO model is dependent upon the recruitment and training qualified educators chosen specifically for their interpersonal skills.[5] They are programmed to provide valuable feedback to their examiner according to their internal sensations and not just to what they can see. As non-medical instructors, they may provide a less threatening learning atmosphere than students often encounter with physician instructors. The non-medical teaching associate, however, has little knowledge of pathology which puts limitations on his ability to answer questions. The PEO model is able to utilize the services of a physician with sufficient medical knowledge to answer a wide range of questions based on broad experience and advanced technical skills. However, the levels of communication skills vary greatly among medical staff.

The IEO system appears to produce students who demonstrate a consistently higher degree of accuracy and thoroughness in performing the examination itself and to provide them with significantly better skills for dealing with patients' physical and emotional discomfort compared to PEO teaching. It is apparent from appendices II and III that the students trained by the IEO model were more likely to provide an overview of the procedures to the patient ($p < 0.0001$), maintain terminology at an appropriate level ($p < 0.0838$) and demonstrate concern for the patient's physical and/or emotional comfort ($p < 0.0001$) than did students trained by the PEO method. In addition, they performed a more thorough examination ($p < 0.0001$) and did it in a fashion that indicated to the evaluator that they were better prepared ($p < 0.0001$). Ninety-seven percent of the IEO students appeared "to know what he/she was doing" compared to only 43.4% of the PEO students. Of the PEO group, 17.4% did not appear to their evaluators to know 'at all' what they were doing, whereas no IEO student was categorized as such.

Over 95% of the IEO students were either moderately or extremely thorough in their examination as compared to only 34.7% of the PEO trainees. Students trained by the IEO method were over 18 times more likely to be judged as being "extremely thorough" than the PEO pupils. Only 5.1% of the IEO students, but as many as 43.5% of the PEO trained students, were felt to be "not at all thorough" in their examination techniques.

One hundred percent of the IEO students were adjudged to have performed the examination in a manner thorough enough to discover genital, rectal, or sexual problems. Only 10% of the PEO group was felt to have been this thorough. Of the physician trained group over 47% did not examine all areas of the prostate, and 70% did not examine the rectum thoroughly in all areas; similar figures for the IEO group were 5.7% and 2.4% respectively.

CONCLUSIONS: Our data suggest: 1) the IEO model provides a significantly more effective learning experience for beginning medical students than the PEO model; 2) students are better prepared in technical skills by the IEO than by the PEO; and 3) the IEO students exhibit significantly better affective behavior skills than PEO students.

Another advantage of the IEO method is the flexible scheduling that is possible since teaching-associates can teach during the day when most physicians are too busy to commit large blocks of time for this purpose. The training of the teaching associate simulated patients requires approximately 10 hours of physician time each year (excluding the initial training period of 20-25 hours) to refresh and refine technical and communication skills, and train new teaching associates. For the other half of the class, taught solely by physicians, approximately 55 hours of physician time were used.

The IEO model provides a reasonable alternative for dealing with the socially sensitive issues which arise when students examine either real patients or each other. All of the sophomore medical students were given the opportunity to evaluate the training systems. Those in the PEO group indicated that they were not comfortable being a simulated patient for their classmates. They were slightly more at ease, but still not comfortable, performing the examination on their classmates. On the other hand, the students were more satisfied with the IEO system and felt that this method prepared them adequately for the future.

The IEO method should be increasingly utilized in future training programs. Peer examination should be avoided, since student discomfort, emotional and physical, does not produce a positive learning atmosphere. Investigation should be made to integrate the positive aspects of the IEO model with the medical expertise offered by the physicians in the PEO technique.

BIBLIOGRAPHY

1. R. H. Kretschmar, Evolution of the Gynecology Teaching Associate: An Educational Specialist. Am. J. Obstet. Gynecol. 131 367-373 (1978).

2. A. Behrens, H.V. Barnes, W.L. Gerber, M. Albanese, S. Matthes, et al., An Effective Model for Teaching Sophomore Medical Students to Perform the Male Genital-Rectal Examination. Proc. 17th Annual Conference on Research in Medical Education (AAMC), Oct. 25-26, 1978, pp. 263-268.

3. A. Behrens, H.V. Barnes, W.L. Gerber, M.A. Albanese, S. Matthes, et al., A Model Program for Teaching Sophomore Medical Students the Essentials of the Male Genital-Rectal Examination. In press, J. Med. Education.

4. The Non-Physician in Medical Education, P.L. Stillman, editor, University of Arizona (monograph), Tucson, 1978; Appendix C, pp. 42-46.

5. Sabers, D.L., Laguna, J.E., Ruggill, J.S. and Stillman, P.L., Preliminary Issues and General Considerations, Ibid, pp. 9-12.

Appendix II: Results of Interpersonal Skills - Ratings of Student Performance

QUESTION	IEO STUDENTS	PEC STUDENTS	CHI SQUARE	COEFFICIENT	SIGNIFICANCE
Was an overview of the procedures in a genital-rectal examination given?					
yes	33 (86.8%)	9 (45.0)	9.485	Phi=.44	p < .0021
no	5 (13.2)	11 (55.0)			
Were explanations of procedures given before performed?					
never	1 (2.6)	13 (61.9)	30.13	Cramer's V=0.71	p < .0001
sometimes	5 (13.2)	3 (14.3)			
usually	10 (26.3)	4 (19.0)			
always	22 (57.9)	1 (4.8)			
Was terminology maintained at an appropriate level?					
not at all	0 (00.0)	3 (15.0)	6.89	Cramer's V=0.34	p < .0754
sometimes	1 (2.6)	1 (5.0)			
usually	5 (13.2)	1 (5.0)			
always	32 (84.2)	15 (75.0)			
Were questions invited at appropriate times?					
not at all	0 (00.0)	10 (45.5)	27.92	Cramer's V=0.68	p < .0001
sometimes	7 (18.9)	7 (31.8)			
usually	12 (31.6)	4 (18.2)			
always	19 (50.0)	1 (4.5)			
Were expectations for physical (dis)comfort given immediately before each procedure?					
not at all	1 (2.6)	15 (68.2)	31.04	Cramer's V=0.73	p < .0001
sometimes	4 (10.5)	2 (9.1)			
usually	11 (28.9)	3 (13.6)			
always	22 (57.9)	2 (9.1)			
Was information about findings provided?					
not at all	2 (5.1)	10 (45.5)	15.02	Cramer's V=0.50	p < .0018
sometimes	4 (10.3)	2 (9.1)			
usually	10 (25.6)	3 (9.1)			
always	23 (59.0)	8 (36.4)			
Was concern for patient's physical and/or emotional comfort demonstrated?					
not at all	1 (2.9)	3 (15.0)	23.43	Cramer's V=0.66	p < .0001
sometimes	0 (00.0)	3 (15.0)			
usually	3 (8.6)	9 (45.5)			
always	31 (88.6)	5 (25.0)			

Note: The number of students does not always total 61 since "uncertain" data was treated as if it were missing.

Figures in parentheses indicate percentages for each model.

Appendix III: Results of Technical Skills - Ratings of Student Performance

QUESTION	IEO STUDENTS	PEO STUDENTS	CHI SQUARE	COEFFICIENT	SIGNIFICANCE
Were procedures performed in a manner which minimized discomfort?					
not at all	0 (00.0)	3 (14.3)	28.28	Cramer's V=.69	p < .0001
sometimes	0 (00.0)	7 (33.3)			
usually	7 (18.4)	7 (33.3)			
always	31 (81.6)	4 (19.0)			
Was the urethra (including meatus) examined thoroughly?					
yes	37 (94.9)	5 (22.7)	30.86	Phi=.748	p < .0001
no	2 (5.1)	17 (77.3)			
Was the penis examined thoroughly?					
yes	37 (94.9)	8 (36.4)	21.95	Phi=.64	p < .0001
no	2 (5.1)	14 (63.6)			
Were the testes examined thoroughly?					
yes	39 (100)	10 (45.5)	23.14	Phi=.66	p < .0001
no	0 (00.0)	12 (54.5)			
Was each epididymis examined thoroughly?					
yes	36 (92.3)	10 (47.6)	12.84	Phi=.50	p < .0003
no	3 (7.7)	11 (52.4)			
Were the spermatic cords examined thoroughly?					
yes	35 (89.7)	11 (52.4)	8.67	Phi=.42	p < .0032
no	4 (10.3)	10 (47.6)			
Were the inguinal canals examined thoroughly?					
yes	37 (97.4)	5 (22.7)	33.5	Phi=.78	p < .0001
no	1 (2.6)	17 (77.3)			
Were all areas of the prostate examined thoroughly?					
yes	33 (94.3)	11 (50.0)	12.64	Phi=.51	p < .0004
no	2 (5.7)	11 (50.0)			
Were all areas of the rectum (i.e. lateral and posterior) examined thoroughly?					
yes	37 (97.4)	6 (30.0)	27.60	Phi=.73	p < .0001
no	1 (2.4)	14 (70.0)			
Was the examination thorough enough to discover any genital, rectal or sexual problems?					
yes	38 (100.0)	2 (10.5)	44.27	Phi=.92	p < .0001
no	0 (00.0)	17 (89.5)			
Did the examiner appear to know what he/she was doing?					
not at all	0 (00.0)	4 (18.2)	33.34	Cramer's V=.74	p < .0001
sometimes	1 (2.6)	9 (40.9)			
usually	5 (12.8)	6 (27.3)			
always	33 (84.6)	3 (13.6)			
Rate the overall thoroughness of the examination:					
not at all	2 (5.1)	10 (45.5)	36.56	Cramer's V=.77	p < .0001
somewhat	0 (00.0)	5 (22.7)			
moderately	6 (15.4)	6 (27.3)			
extremely	31 (79.5)	1 (4.5)			

Note: The number of students does not always total 61 since "uncertain" data was treated as if it were missing.

Figures in parentheses indicate percentages for each model.

PRECIS

ASSESSMENT OF CLINICAL REASONING

MODERATOR: Carl F. Hinz, M.D.
University of Connecticut

THE UNDERLYING STRUCTURE OF CLINICAL PROBLEM-SOLVING: PROCESS OR CONTENT?

Seventy-one medical students who had completed the first two years of a medical program designed around body-system courses were given an examination designed to measure the clinical problem-solving process. Gathering data on patient history and hypotheses generation were found to be specific skills common to the solution of clinical problems while hypotheses refinement, physical examination and laboratory investigation were found to be problem related.

THE "TRIPLE JUMP" EXERCISE - A STRUCTURED MEASURE OF PROBLEM SOLVING AND SELF DIRECTED LEARNING

The paper describes a new evaluation method, the "Triple Jump", designed to assess problem-solving, knowledge, and self-directed learning using a structured oral examination format.

EVALUATING ANALYTICAL REASONING IN THE MANAGEMENT OF THE MULTIPROBLEM PATIENT: A SYSTEM FOR ATTAINING RELIABILITY AND VALIDITY

A systematic approach to describing errors in diagnostic reasoning allows raters to evaluate student performance with surprising interrater reliability. Early experiments with model cases have now been extended to the analysis of real patient records with similar reliabilities.

THE UNDERLYING STRUCTURE OF CLINICAL PROBLEM-SOLVING: PROCESS OR CONTENT?

Harasym, P.H.[1], Alexander, F., Baumber, J.S., Bryant, H., Fundytus, D., MacPhail, I., Preshaw, R., Sosnowski, M., Watanabe, M., Wyse, G. Faculty of Medicine, University of Calgary.

Central to the effective delivery of health care, is the physician's ability to master the complex skill of clinical problem solving. It is therefore vital that medical educators and practitioners develop a broader understanding of this complex skill in order to help students become effective medical problem solvers. Currently, a great deal of confusion exists regarding the extent to which medical content and problem-solving ability underlie the solution of clinical problems. Four theories exist. The first theory advances the hypothesis that clinical problem-solving is case specific and that no common problem-solving process exists. The second theory posits that the process of data gathering is common to all cases and that management is unique to each case. The third theory suggests that the process of data gathering and management underlies the solution of clinical cases. The fourth theory states that four processes, namely, formulation of a hypotheses, selection of correct laboratory procedures, avoidance of incorrect laboratory procedures and formulation of final diagnosis underlie clinical problem-solving. The purpose of this paper is to provide supporting evidence of a possible fifth theory.

Literature Review

The following studies support the theory that clinical problem-solving is case specific. Sprafka and Elstein (1974) used three role playing simulations to develop a model of medical problem-solving and found that performances differed among clinical problems. They concluded that the physicians' performance was case specific. McGuire (1976) interpreted low correlations between total scores on branching patient management problems to indicate that problem-solving skills were unique to each case. Elstein, Shulman, and Sprafka (1978) found that both physicians' and medical students' diagnostic effectiveness varied according to the nature of the medical problem at hand. They concluded that clinical competence may be case related.

Donnelly, Gallagher, Hess and Hogan (1974) found one process underlies clinical problem-solving. They performed a factor analysis of the correlation matrix derived from students' scores on each of ten branching patient-management problems. Two factors consistently described variation in performance: 1) an information gathering factor composed of history, physical, laboratory and X-ray variables; and 2) a decision making factor composed of variables such as diagnoses and management interventions. The authors argued that, since two factors emerged, the unitary score correlated and analyzed in other studies was not a true reflection of the data. When the authors factor analyzed all variables across all ten cases they again found a single information-gathering factor, but they did not find a general factor for decision making. They concluded that information-gathering ability may generalize from case to case, but decision making ability (both diagnostic and management) may be specific to the content of the case.

The following studies found data gathering and management processes to underlie clinical problem solving. Juul, Noe and Nerenberg (1977) administered twenty-four patient management problems to both senior and junior medical students. A score for history, physical, diagnostic procedures, pathway, treatment, and diagnoses was obtained for each student by summing the assigned weights for options across the twenty-four problems. The authors found that the scores for history, physical and diagnostic procedures loaded onto one factor, which they refer to as "data gathering". The scores for pathway, treatment, diagnosis, and diagnostic procedures loaded onto the second factor which they referred to as "management". A similar study by Skakun (1978) analyzed the results of five linear patient management problems which were administered to 103, fourth year medical students. Two data sets were generated. The first, consisted of a history, physical, laboratory, and management score for each student on each problem. The second consisted of a history, physical, laboratory and management score derived by summing across all five problems. An analysis of the first data set produced nine factors, a finding that was at variance with those of Donnelly et al. (1974). Analysis of the second data set, however, produced the two factors of "data gathering" and "management".

---

[1] Direct correspondence to: Peter Harasym, Division of Educational Planning and Assessment, The Faculty of Medicine, 2920 - 24th Avenue, N.W., Calgary, Alberta T2N 1N4.

A study by Berner, Bligh and Guerin (1977) found four processes underlying clinical problem solving. Berner et al. (1977) argued that many of the previous studies did not employ methods which allowed process and content aspects to be separated. They administered five examinations to 190 second year medical students. The students were required to write a problem list, select appropriate laboratory investigations, and identify the final diagnosis. After each task, corrective feedback was given so that all students had the same information at the beginning of the next task. The results derived from factor analysis revealed that the same tasks, irrespective of the clinical problem, clustered together. That is, formulating a problem list loaded onto factor I, selecting correct laboratory procedures loaded onto II, avoiding incorrect laboratory procedures loaded onto III, and making the final diagnosis loaded onto IV. Since this structure of loadings tended to occur for the five cases, the authors concluded that problem solving skills do exist and that they vary among individuals.

The above studies have several shortcomings. For example, Berner et al. (1977) used a procedure which enabled process to separate from content but which did not model the clinical problem-solving process. Their findings therefore may be a product of the "low fidelity" setting and not a reflection of problem-solving in reality. The design used in the patient management problem studies of Skakun (1978), Juul et al. (1977), McGuire (1976) and Donnelly (1974), although being of "moderate fidelity", may have restricted the separation of process from content. The studies described by Elstein et al. (1978) have added much to our understanding of clinical problem-solving; however, their studies are based upon indepth investigations of small samples of subjects. The finding of case specificity could be a product of the educations of the physicians investigated.

The study reported in this paper attempted to overcome some of the above short-comings. It is structured on the premise that if clinical problem-solving is highly dependent on specific medical knowledge in each case, correlations between different skills within a case will be higher than the correlations between similar skills across cases. In other words, generating hypotheses, gathering patient data on history, identifying relevant physical examinations, ordering laboratory investigations and making a final diagnosis would correlate highly among themselves if problem-solving is dependent upon the specific medical knowledge in each case. Generating hypotheses in one case would not be expected to correlate highly with generating hypotheses in another case.

Method

In the spring of 1979 an examination designed to measure the clinical reasoning process was administered to 71 medical students who had completed the first two years of a medical program designed around body-system courses. The examination consisted of three simulated clinical problems which tested students' ability to integrate knowledge acquired from the various courses. The problems varied both in the nature and the urgency of management.

The first clinical problem presented a 46 year old woman with abdominal pain, the second presented a one year old male infant of short stature, and the third presented a 64 year old man who suddenly sank to the floor unconscious. The practicioner's office was selected as the setting for the first two problems and the main lobby of a hospital for the second.

In these three problems, the student proceeded in sequence from one task to the next, and handing in the task of each station before proceeding to the next. The first two problems were divided into seven stations and the third into five. The format of problems 1 and 2 are given in Table 1.

Table 1: Structure of Integrative Examination. Problem 1 and 2.

Given	Student Response	Station
brief history	gives hypotheses list 1	I
simulated patient		II
	gives hypotheses list 2	III
	identifies historical findings	IV
complete history	gives hypotheses list 3	V
	identifies relevant physical findings	
complete physical findings	gives hypotheses list 4	VI
	orders lab. investigations	
complete lab. results	makes final diagnosis - hypothesis 5	VII

In station I, a student was given a brief history and asked to give an initial set of hypotheses. The history stated:

> A 46 year old woman whom you have never seen before arranges an urgent appointment to see you in your family practice office. She tells your nurse that she has had abdominal pain for the last two months which is getting worse, and she has also lost 15 pounds in weight. List five body systems which might develop a process presenting in this fashion and give one example in each system.

The list, known as hypotheses list 1, was collected by the invigilator before the student proceeded to the next station. In station II, the student was given ten minutes with a simulated patient to obtain whatever historical information was considered relevant. The student had been advised to use selective questioning since insufficient time was available to take a complete history. Each student was allowed to make notes and keep these for the duration of the problem. Upon obtaining a history, the student proceeded to stations III - VII.

In stations III - VII the student was asked to complete the following tasks:

(i) station III, list a differential diagnoses based upon the obtained history (i.e. hypotheses list 2);
(ii) station IV, answer a multiple-choice test on historical findings;
(iii) station V, refine the differential diagnoses using a complete, accurate history (i.e. hypotheses list 3); in addition, list specific physical investigations to be performed;
(iv) station VI, refine the differential diagnoses using complete accurate history and physical findings (i.e. hypotheses list 4); in addition, order the required laboratory tests;
(v) station VII, make a final diagnosis based upon complete, accurate history, physical plus laboratory findings (i.e. hypotheses list 5).

Problems 1 and 2 required eight written student responses: five of either hypotheses generation or refinement, and three of gathering historical information, identifying relevant physical examinations and ordering laboratory investigations. Corrective feedback was provided after each of the latter three tasks.

The third problem was an emergency situation of an unconscious patient. Its format was altered to reflect a normal sequence of events that might occur in management of the patient. The problem format is summarized in Table 2.

Table 2: Structure of Integrative Examination. Problem 3.

Given	Student Response	Station
brief history	lists physical signs	I
results of physical signs	gives hypotheses list 1 identifies relevant physical findings	II
complete history and physical examination	lists hypotheses list 2 orders initial lab. investigations	III
initial lab. results	orders remainder of lab. investigation	IV
additional lab. results	final diagnoses - hypotheses list 3	V

In station I, the student was given a brief description of the setting and presenting problem. For example:

> You see a 64 year old man enter the main lobby of the Foothills Hospital, walk briskly towards the elevators, stop and suddenly sink to the floor unconscious. You quickly go to him.

The student was then asked to list the four specific signs that should be examined for immediately. The patient is then transferred to the emergency ward. In the next station, the specific signs are given and the student was asked to generate a differential diagnoses (i.e. hypotheses list 1) and to list a physical examination to be performed. In station III a complete, accurate physical examination and history were provided and the student was instructed to revise his/her differential diagnoses (i.e. hypotheses list 2) and to order the pertinent laboratory investigations. In addition, the student was reminded that the patient was still in the emergency room. In station IV the student received the results of the initial laboratory investigations, was told the patient was admitted to hospital, and was asked to order the remaining laboratory investigations. In station V the results of the additional laboratory investigations were provided and the student was directed to make a final diagnoses (i.e. hypotheses list 3).

Twenty-three scores were generated for each student, one for each of the student responses outlined in Tables 1 and 2. Each score reflected the degree each task was correctly completed.

Principal component analysis was performed on the matrix of correlations coefficients among the twenty-three scores. Components with eiginvalues greater than one were retained and transformed by the Harris-Kaiser transformation procedure.

Results

The results of the rotated components are presented in Table 3. The component loadings falling outside the range of ±0.35 are reported. The four components accounted for 65.7% of the observed variance. The variable labels of HYP, HIST, PHY and LAB respectively represent generating and/or refining hypotheses, gathering data on patient history, identifying physical examinations, and ordering laboratory investigations. The number following the label identifies similar tasks with the same problem.

Based upon their loadings, Component I was referred to as 'data gathering and interpretation', Component II as 'hypotheses generation', Component III as 'medical knowledge specific to the problem', and Component IV as 'hypotheses restructuring based upon corrective feedback'. The first three components are believed to underlie clinical problem-solving and the fourth to be an artifact of the experimental design.

Discussion

The findings of this investigation are at variance with other studies. Firstly, they do not support the concept that clinical problem-solving is case specific. They do however lend support to those theories which postulate that general problem-solving abilities do exist. In this study, it was observed that gathering data on patient history and generating hypotheses loaded onto different components irrespective of the clinical problems. It was also observed that hypotheses refinement; physical examinations, and laboratory investigation could load on the same or on separate components. The reason for this variation was not apparent, but it could be due to the nature of the problems or the experience and training of the subjects. The former cause is believed to be more probable, but this idea requires further investigation.

In summary, this investigation provided evidence to suggest that gathering data on patient history and hypotheses generation are specific problem-solving skills common to the solutions of clinical problems, while hypotheses refinement, physical examinations and laboratory investigations are problem related. Further investigtions in this area are needed in order to broaden our understanding of these complex skills.

Table 3. Principal Component Analysis of Three Simulated Clinical Problems: Oblique Factor Pattern Matrix.

Variable	Component				
	I	II	III	IV	Communality
HYP1				-.39	.15
HYP2		.84			.75
HIST	.85				.83
HYP3				-.91	.94
PHY1					.15
HYP4	.87				.89
LAB1		.78			.70
HYP5	.71				.60
HYP1			.52		.44
HYP2		.83			.77
HIST	.76				.58
HYP3				-.69	.83
PHY1	.76				.67
HYP4	.86				.89
LAB1			.58		.54
HYP5		.84			.72
PHY1			.36		.35
HYP1		.78			.80
PHY2			.41	-.43	.75
HYP2	.82				.65
LAB1	.88				.86
LAB2	.93				.85
HYP3			.50		.40

REFERENCES

Berner, E.S., Bligh, T.J., and Guerin, R.O. An indication for a Process Dimension in Medical Problem Solving, Medical Education, vol. 11, no. 5, pp. 324-328, 1977.

Donnelly, M.B., Gallagher, R.E., Hess, J.W., and Hogan, M.J. The Dimensionality of Measures Derived from Complex Clinical Simulations. Proceedings, Conference on Research in Medical Education, Association of American Medical Colleges, 1974.

Elstein, A.S., Shulman, L.S., and Sprafka, S.A. Medical Problem Solving. Cambridge, Harvard University Press, 1978.

Juul, D.H., Noe, M.J., and Nerenberg, R.L. A Factor Analytic Study of Branching Patient Management Problems. Paper presented at the annual meeting of the National Council of Measurement in Education, New York City, April, 1977.

McGuire, C. Simulation Technique in the Teaching and Testing of Problem Solving Skills. Journal of Research in Science Teaching, 13(2), p. 89, 1976.

Skakun, E.N. The Dimensionality of Linear Patient Management Problems. Proceedings of the Seventeenth Annual Conference on Research in Medical Education, New Orleans, Louisiana, October, 1978, 63-67.

Sprafka, S.A. and Elstein, A.S. What Do Physicians Do? An Analysis of Diagnostic Reasoning. Paper presented at Conference on Research in Medical Education held in conjunction with annual meeting of The Association of American Medical Colleges, November, 1974.

THE "TRIPLE JUMP" EXERCISE - A STRUCTURED MEASURE OF PROBLEM SOLVING AND SELF DIRECTED LEARNING

C. Painvin, V. Neufeld, G. Norman, I. Walker, G. Whelan
Program for Educational Development
McMaster University, Hamilton, Canada

The undergraduate M.D. Program at McMaster University emphasizes problem based learning, self-directed learning, small group tutorial learning and ongoing evaluation(1). More specifically, the student evaluation system is characterized by the following features:

(1) Attention is paid to a wide range of learning objectives.
(2) Measures are chosen or designed which attempt to match a particular objective.
(3) Self, peer and tutor assessments are used selectively.
(4) There is an emphasis on feedback to the student as the primary purpose of the evaluation; the same observations are used for both "diagnostic" and "certification" purposes. There are no end of unit or course examinations.

While many of the objectives of student performance are based on informal observations in the tutorial setting, it became apparent that these were insufficient. It was difficult in the group setting to be sure which contributions of the student were independent of the contributions of peers. The climate within the tutorial is informal where learning and evaluation are intermixed, and where the role of the tutor is primarily that of a facilitator. Furthermore, some tutors might be relatively "nonexpert" with respect to some of the content being discussed. It became apparent that an individual measure of problem solving ability was required.

A number of measures of student performance were considered. Multiple choice format focuses primarily on the gain of factual information and does not focus clearly on problem solving ability and the application of information to problem situations. Patient management problems are difficult to construct and have low validity in terms of their relationship to performance with actual patients(2). Direct observation of a student-patient encounter is more difficult to arrange logistically, focuses on a broader range of skills including clinical data acquisition, and at the early stages of our program was thought to be premature. A written analysis of the synthesis of a problem outside of an examination room setting was found to be difficult to standardize.

Given these considerations, a three-step measure of individual problem solving was devised which was nicknamed the "triple jump" exercise.

Description of the Triple Jump Exercise

There are three steps:

Step 1 - Initial analysis of the problem by the student in the presence of an

Send reprint requests to Dr. C. Painvin at the above address.

94

assessor. This includes hypothesis generation, acquisition and interpretation of clinical evidence, and an initial exploration of underlying mechanisms.

Step 2 - Information search independently by the student. This is usually over a limited period of time such as 2 hours.

Step 3 - Synthesis: This again is in the presence of the assessor. Here the student describes the process of his information search (how personal objectives were set and prioritized, which resources were used and how the time was spent). He then presents his final analysis of the problem, and the highlights of the new knowledge obtained as it relates to his understanding of the problem. The student is given an opportunity to assess his overall performance in the exercise, and this is followed by feedback by the assessor to the student.

In the form used in the present study, the triple jump consists of a written case protocol containing features of the presenting problem (history, physical and investigations), a criterion list of diagnostic hypotheses the student should consider in Step 1, a criterion list of suggested issues for exploration in Step 2, and fourteen categories of evaluation, interspersed through the written protocol in the sequence they would be encountered as the student works through the problem.

The triple jump exercise is thought to have the following advantages:

1. It is based on systematic research into the nature of clinical reasoning in physicians and medical students(3).
2. It incorporates several of the general objectives of the M.D. Program:
   (a) understanding of underlying mechanisms.
   (b) problem solving (clinical reasoning) ability.
   (c) self-directed learning ability.
   (d) self appraisal skills.
3. The exercise is very similar to how students actually learn in the program.
4. There is increased objectivity of assessment by standardizing the format, the "data base" of the problem, and the criteria of student performance.
5. There is opportunity for direct feedback to students built into the exercise.
6. The exercise is flexible and can be adapted to various situations. For example:
   (a) either a paper or a problem card deck (P4) format can be used for the initial problem presentations.
   (b) the tutor himself may be the assessor, or an "external" assessor may be recruited for the exercise.
   (c) the final synthesis can be verbal or written. Some planners have used a "consultation to the referring physician" as the form of the final synthesis.

In order to determine the value of the triple jump exercise a small study was initiated in the Phase III (Year II) hematology-gastroenterology unit. This unit is held four times in the academic year, with twenty-five students in each session. The study was undertaken in the spring session, involving a sample of twenty-five students.

The objectives of the study were:

(1) Assessment of inter-observer reliability of the technique.
(2) Assessment of generalizability of the technique across different problems.
(3) A measure of the validity of the technique by comparison with tutor evaluation.
(4) Assessment of the acceptability of the technique to students and tutors.

Experimental Method

The twenty-five students involved in the study were introduced to the study methods at an informal meeting prior to the first evaluation.

Two evaluation exercises were used in the study - the first was conducted halfway through the ten week unit using a hematology problem, and the second was conducted at the end of the unit with a gastroenterology problem. The evaluators for the exercise were tutors participating in the unit; at the mid-term exercise tutors "swapped" groups, and at the final exercise tutors evaluated students from their own group.

Because of the concern that students within a tutorial group may discuss problems, five different problems were developed for each exercise and randomly assigned to students in each group.

A subsample of two groups (10 students) was observed by one of the investigators (C. Painvin or G. Whelan) and independently rated on the same form to assess inter-observer reliability.

Finally a form containing the categories of evaluation on the triple jump protocol was sent to tutors just prior to the mid-unit evaluation, and they were asked to assess students in their groups on the identified categories to be used as a criterion measure for the assessment of validity.

A form designed to assess satisfaction with the triple jump exercise was also sent to students and tutors who were involved in the unit over the past year (99 students and 20 tutors).

Results

Correlational analysis was used to assess the reliability and validity of the evaluation data. The results are summarized in Table 1.

The first two columns tabulate inter-observer reliability for the subsample of ten students on the mid-term and final evaluation. For the mid-term evaluation, correlations are reasonably high with the exception of categories 1 and 2, which relate to early diagnostic hypotheses. As a result of this low reliability, criterion hypotheses were introduced in the written protocol for the final evaluation, and resulted in a substantial improvement in reliability. Looking in detail at the correlations for the final evaluation, categories related to self-assessment and self-directed learning are extremely low, although these were adequate at mid-term. Two changes in the method may explain this reduction; in the final evaluation, students were given 24-48 hours for self-directed learning, in contrast with

2 hours at mid-term. Apparently this skill cannot be adequately assessed unless time constraints are imposed. Also, in the final evaluation, tutors were assessing their own students, which may have contaminated their assessment of self-assessment ability, leading to poor agreement with the observer.

The third column indicates the correlation of performance from the mid-term to final evaluation. With the exception of the categories of self-directed learning, which were unreliable, the correlations range from .15 → .50, comparing favourably with correlation of performance across problems for objective measures of problem-solving such as the P.M.P.

The final two columns are an attempt to assess the validity of the method by correlation with tutor evaluation. Evidently there is little relationship between the two methods. Whether this is due to low validity of the triple jump or a poor choice of criterion measure can only be speculated. However the criterion measure may be inadequate, as standard deviations of tutor evaluations were in the range 1.2 → 1.5 on a ten-point scale, approximately half those observed on the triple jump, and suggesting a lack of discrimination. Average scores on the tutor report were also high, ranging from 7.1 → 7.9 on the ten-point scale.

The satisfaction questionnaire was sent to 99 students and 20 tutors who had been involved in the unit over the past year. Response rate was 90% from tutors and 75% from students. The questions asked subjects to estimate the relative value of the triple jump and the traditional evaluation of written problem write-ups. Eighty-three percent favoured the triple jump, 8% the problem-write-up, and 9% had no preference. Similarly, 87% were positive in their opinion of the value of the triple jump.

Discussion and Conclusions

The triple jump exercise was introduced into the McMaster program as an attempt to improve on existing tutorial-based evaluation activities, without compromising the opportunity for feedback to students or suffering from many of the deficits of existing evaluation methods.

The present study has demonstrated acceptable inter-observer reliability of the format, and very positive acceptance by students and tutors. The validity of the method has not been confirmed, possibly because of the absence of an adequate criterion measure in the McMaster program. In addition it remains to be determined whether the format can be usefully applied in other institutions, or whether the format has found acceptability at McMaster because of the unique aspects of the educational system.

References

(1) J. D. Hamilton, "The McMaster Curriculum: A Critique", I 1191-1196, 1976.

(2) J. W. Feightner and G. R. Norman, "Concurrent Validity of Patient Management Problems by Direct Comparison with the Clinical Encounter", Proc. 15th Conference on Research in Medical Education, San Francisco, November, 1976.

(3) H. S. Barrows, V. R. Neufeld, J. W. Feightner and G. R. Norman, "Studies of the Clinical Reasoning Process of Medical Students and Physicians", Proc. 16th Conference on Research in Medical Education, Washington, November, 1977.

Table 1

Reliability and Validity Coefficients of the Triple Jump Evaluation

	Reliability		Correlation	Validity	
	Mid-	End	across Problems	Mid-	End
Sample Size	8	8	24	24	19
Hypothesis Generation					
Demonstrates understanding of key mechanisms involved in problem.	.34	.77	.39	-.18	-.30
Hypotheses consider pertinent etiologies and their prevalence, are appropriately specific.	.40	.90	.50	-.07	-.23
Data-gathering: Appropriate amount of data gathered. Critical positive and negative findings elicited and correctly interpreted.	.50	.64	.15	-.01	.14
Problem Formulation: Final problem formulation adequately characterized patients problem(s). Likely diagnostic alternative considered.	.45	.60	.34	-.02	.21
Investigations: Investigations are appropriate. Consideration of cost/risk and predictive value.	.87	.51	.31	.04	.18
Issue Generation					
Issues raised by student match critical issues in problem.	.29	.25	.27	-.08	.74
Self-Assessment: Student identifies issues related to won areas of deficiency.	.86	.08	.15	-.21	.28
Self-directed Learning					
Pursued issues and established priority for study.	.76	0	.05	-.02	.00
Evidence that appropriate resources consulted.	.76	.23	.01	.51	.01
Efficient use of time.	.51	.26	.02	.07	-.18
Synthesis					
Demonstrated critical thinking and conceptual understanding.	.84	.64	.26	.14	.21
State of Knowledge					
Student's knowledge and understanding as demonstrated at completion of the problem is appropriate for this unit/phase.	.71	.59	.22	.03	.28
Self-Assessment					
On direct questioning, student adequately identified strengths/weaknesses.	.67	.37	.23	-.34	.18

EVALUATING ANALYTIC REASONING IN THE MANAGEMENT OF THE MULTIPROBLEM PATIENT: A SYSTEM FOR ATTAINING RELIABILITY AND VALIDITY

A. E. Voytovich, M.D., L. Harper, M.D., R. M. Rippey, Ph.D.,
University of Connecticut Health Center, School of Medicine,
Farmington, CT 06032*

Purpose

Current techniques for measuring clinical performance are many and diverse; however, there is one critical clinical task for which no adequate proven method of measurement has been useful. That task is the assembly of multiple clinical clues into a working set of hypotheses and diagnoses in the patient with multiple problems.

"Patient-management problems," frequently focus on a single unifying diagnosis and stress data gathering. Simulations using paper and pencil, actors, or computers frequently also focus on development of a single diagnosis. Many medical patients have more than one diagnosis or problem, most have five, many even more(1). In addition, the problems include a mixture of defined etiologic entities, undefined pathophysiologic states and isolated abnormalities all interacting to produce a constantly changing picture.

Not only have available methods of measurement avoided the multiproblem patient, they have also failed to study uncertainty -- where uncertainty exists, and usually require at least a diagnostic conclusion as an end point. Furthermore, traditional measures of clinical performance (exams, patient management problems, simulations) have not led directly to evaluation techniques which might be applied to real clinical settings.

We have experimented with a technique for measuring the ability to synthesize the multiple clues of the complex patient into etiologic diagnoses, working hypotheses and isolated abnormalities -- with credit for acknowledging uncertainty where uncertainty exists. The reliability and validity of the technique have been demonstrated using model cases. In this study we report on the extension and testing of our observations into the analysis of real cases.

In earlier studies with model cases (2,3), we showed that when the student tries to produce an accurate, precise problem list, a limited number of reproducible errors in judgment occur. These errors can be characterized to the point that a quantitative estimate is possible. Interrater reliability is excellent and the number and type of error found correlate with clinical experience to a high degree(4). The present results show that the analytic framework emerging from these "scoreable problem lists" can be employed to audit real cases with similar reliability.

---

*Address reprint requests to Dr. Voytovich

Background

Elstein and Shulman (5,6), led the way in their studies of diagnostic reasoning. There seems to be general agreement that physicians generate diagnostic hypotheses early in the clinical interaction, and proceed to test these, rank ordering them by probability, seriousness, treatability and novelty. The number of diagnostic hypotheses varies from two to six. Considering this scheme in the context of prior understanding of human information processing, the clinician rapidly encounters difficulty with the magnitude of his task. Simon (7) suggested that disorganized inputs of large amounts of information can quickly overload a person's processing capacity. He has developed the notion of the "chunk" as a unit of information which can be moved rapidly between long and short-term memory; Elstein and Shulman (5) suggested that early hypothesis formation is a form of chunking designed to handle the rapid accumulation of single facts or clues.

Not only are the number of provisional hypotheses limited by short-term memory, but rank ordering by probability is open to error. Tversky and Kahnemann (8) described the common and systematic misconceptions inherent in subjective probability assessment. Certain apparently effective and economical strategies in judgment in the presence of uncertainty lead regularly to error. These biases include "insensitivity to prior probability and sample size, misconceptions of chance and regression, retrievability of instances and effectiveness of a search set," as well as "anchoring and adjustment." These biases operate regularly. The resident says, "This can't be Wilson's disease we just had one last week," ("misconception of chance"). Elstein and Shulman (6) recognized "anchoring and adjustment" -- a tendency to undervalue new information which is a phenomenon common to students of Bayesian analysis.

We have characterized the assembly of clinical clues into an accurate problem list as similar to the sorting and assembly of a pile of jig saw pieces with additional provisions that: a) there may be more than one puzzle; b) any one or all may be incomplete; c) some pieces fit more than one of the puzzles; d) some pieces don't fit any of the puzzles.

The problem list can act as an accounting system to keep track of the completed, partially completed, and fragmentary clues all of which reflect the clinician's understanding of conclusions consistent with available data.

Analysis of over a thousand records during the past 10 years has shown that subjects fail to recognize all the abnormal data and draw faulty conclusions in a surprisingly consistent way. All errors can be classified into one of the following four types:

1) OMISSION - A critical abnormality is not considered in the analysis.

2) PREMATURE CLOSURE - A conclusion goes beyond the available facts.

100

3) INADEQUATE SYNTHESIS - A pathophysiologic state or diagnosis is not recognized even though sufficient data are present to allow the conclusion.

4) WRONG SYNTHESIS - The data clearly contradict the conclusions.

The fact that the errors always seem to conform to one of these types has had great utility. We have described the relative frequency of these phenomena and drawn conclusions concerning performance. A count of errors on problem lists suggests that students who were felt by their attendings to be weak made more of the above errors and better students fewer(3). Of course it is no surprise that poor students don't do well, but this error analysis shows promise as a quantitative as well as qualitative measure.

Experiments were initially conducted using model cases to determine whether problem lists could be scored with acceptable interrater reliability, whether there was any improvement in student performance with experience, and whether parallel case reliability could be shown.

Using model cases, subjects were asked to construct problem lists de-novo in a completely open system with no cueing. Using a simple scoring key and a common appreciation of the four errors, interrater reliability greater than .85 was attained, even when scorers were not physicians(2).

Using two similar cases, correlations over parallel forms were in the range of .55. Using the Spearman Brown prophecy formula (9), this would represent reliability of approximately .85 for a hypothetical ten-item test. Thus, performance did not seem to be case-specific and individual behavior seemed consistent.

The question remained: Could the concept of the four errors and the surprising interrater reliability be extended to audit of real cases?

Methods

A series of 50 patient records were selected at random from the inpatient service of the University of Connecticut John Dempsey Hospital. Each case was rendered anonymous by deletion of names, alteration of occupation, age or sex. The data were abstracted into two typed pages and included history, physical examination, laboratory findings and hospital course. The actual problem list was copied. No attempt was made to clarify issues; all the ambiguity and background confusion of the patient with multiple problems were preserved.

Precise instructions for scoring were generated, specifying the four errors and directing three auditors to independently score each problem list, counting the number of instances of each error. Auditors A and B met after each eight or ten cases to compare scores, and determine the need for improving the instructions. Auditor C operated independently. Auditors spent about 15 minutes per case and recorded the number and type of errors on each problem list.

The resulting counts of errors were compared to determine interrater reliability using the PLATO computer system statistics package. Correlation coefficients were corrected using the Spearman Brown prophecy formula. In addition, correlations among the variables were calculated to determine whether the four errors were independent or tended to occur together.

Results

Utility of the four errors -- Wrong synthesis occurred so infrequently as to make any analysis insignificant. The remaining three types of errors, in all cases, proved sufficient to describe each bad synthesis on the 50 problem lists. It became clear, however, that "inadequate synthesis" fell into two distinct varieties: incomplete specification (such as referring to rheumatic heart disease with mitral stenosis and CHF as "heart disease") and "failure to combine" (such as referring to edema, rales, cardiomegaly gallop rhythm as separate entities rather than as CHF).

Frequency of occurrence of the four errors -- Mean number of errors per case:

Auditor	Omission	Incomplete Synthesis	Premature Closure	Wrong Synthesis
A	2.94	1.06	.61	.02
B	3.49	.96	.71	.07
C	2.84	1.18	.65	.02

Omissions were most frequently found, wrong synthesis rare, and the others ranged between. This relative distribution is similar to that reported earlier in a single auditor's experience at a different institution (3).

Interrater reliability -- The numeric error profiles generated by each auditor were quite similar considering the diversity of the cases and the fact that the subject responses were completely open, unbounded and uncued. Correlations between the three auditors were determined for each error type:

Rater Pair	Omission	Incomplete Synthesis	Premature Closure
A:B	.89	.83	.94
A:C	.79	.77	.85
B:C	.84	.69	.87

All correlations are significant at the $p < .00005$ level. Recall that Auditor C worked independently using only written instructions. It would appear that excellent interrater agreement can be achieved using the written instructions alone.

Correlations among variables -- Correlations among the error types, case by case were very low (r = between -.005 and .32). This shows that the different errors represent distinct entities and are not simply manifestations of some underlying global ability.

Discussion -- Conclusions

The utility of considering errors in problem formulation within a defined framework has been established. The four errors (omission, inadequate synthesis, premature closure, and wrong synthesis), have proven sufficient to describe any faulty synthesis in a problem list. Using this concept of description, excellent interrater reliability has been demonstrated, first with simulated model cases, and now with real cases. A minimum of training is required to prepare an auditor.

The variables (errors) appear to be independent, thus measure different skills in synthesis.

Early studies in a structured setting using model cases revealed dramatic improvement in performance when second, third and fourth-year students, residents and faculty were compared. The groups are clearly separable not only in the number, but by the type of error committed. Clearly the ability to recognize and correctly specify complex entities contributed most to the difference(4). Experiments are currently underway to determine whether improvement in problem definition occurs as a function of advanced clinical training, and whether the differences observed in the structured setting can be found when auditing real cases.

This technique would appear to have promise in the quantitative as well as qualitative measurement of analytic ability in the audit of real cases. It is more practical and realistic than those techniques which require closure on a single diagnostic entity. The task in this method is to sort and sift the many clues in the complex, multiproblem patient -- acknowledging uncertainty where appropriate and conclusions where they are justified.

REFERENCES

(1) Froom, J. Conversion to Problem Oriented Records in an Established Practice. *Annals of Internal Medicine* 78:254, 1973.

(2) Voytovich, A., Rippey, R. Audit of the Structured Problem List as a Measure of Clinical Judgment: Reliability and Validity. *J of the Irish Med. Assoc*. 78:346, 1978.

(3) Voytovich, A. An Analysis of Student Performance in the Non-Threatening Audit of the Structured Medical Record. *Academic Decision Making, Issues and Evidence*. AAMC Council of Deans, pp 37-42, 1975.

(4) Voytovich, A., Rippey, R., Copertino, L. Structured Problem Lists as Measures of Clinical Judgment, (submitted for publication).

(5) Elstein, A., et al. Methods in the Study of Medical Inquiry. *J Med.Ed.* 47:85-92, 1972.

(6) Elstein, A., et al. Medical Problem Solving, An Analysis of Clinical Reasoning. Harvard University Press, Cambridge, 1978.

(7) Simon, H. A. How Big is a Chunk? Science, 183:482, 1974.

(8) Tversky, A., Kahnemann, D. Judgment Under Uncertainty. Heuristics & Biases, Science 185:1124, 1974.

(9) Gulliksen, H. Theory of Mental Tests, New York, John Wiley & Sons, 1962.

The authors thank Carl F. Hinz, M.D. for his patience in serving as Auditor C and for his critical review of this manuscript.

PRECIS

PROFESSIONAL DEVELOPMENT

MODERATOR: H. Tristram Engelhardt, Jr., M.D., Ph.D.
Georgetown University

ASSESSING THE MORAL DEVELOPMENT OF MEDICAL STUDENTS: AN EMPIRICAL STUDY

This study provides evidence which demonstrates that the moral judgement of medical students continues to develop during their professional studies. Development appears to be a function of interpersonal relationship style, which may have implications for the development of medical humanities curricula.

MORAL REASONING AND PHYSICIANS' DECISIONS IN CASES OF CRITICAL ILLNESS

A study of the relationship between a physician's stage of moral reasoning and the degree of activism with which he or she treats cases of critical illness.

COPING AND EGO DEVELOPMENT IN INTERNSHIP

This paper reports on an in-depth study of personal and professional development during internship. Stages of ego development and self-reported coping mechanisms are utilized to examine interns' self-assessments of their experience and faculty ratings of clinical performance.

Assessing the Moral Development of Medical Students: An Empirical Study.

M. Harry Daniels, Ph.D., Assistant Professor, Guidance & Educational Psychology, Southern Illinois University at Carbondale. George L. Baker, M.D., Professor and Associate Dean, College of Medicine, University of Iowa.

Purpose

Recent advances in the basic sciences have brought about numerous moral and ethical questions for the medical profession. The profession has responded to this pressing issue in a variety of ways. One response has been to include courses in humanities and ethics in the medical curricula. Despite the demonstrated interest in such courses, the dearth of information about the moral/ethical development of medical students has limited the ability of concerned educators to make informed decisions concerning the structure and content of such courses. Given these parameters, there is an obvious need for empirical evidence that will provide decision makers with some much needed information. The results of the present study will supply some of that data.

This study represents a longitudinal analysis of the moral development of medical students as they progress through their medical studies. Two questions were framed to guide the research efforts. First, do medical students exhibit advances in moral judgment during their enrollment in medical school? Second, are there differences in moral judgment development among groups of students with different styles of interpersonal participation?

Literature Review

Medical schools, like other educational institutions, have been charged with the responsibility of socializing the students in attendance (1). But unlike most other educational institutions, medical institutions have a dual task. In addition to fostering the general socialization process, medical educators must socialize the students into a select professional role.

In the last half century, the major emphasis of medical education has been on the acquisition of technical knowledge, with less attention directed toward the "non-intellectual components" (2) of medical education. Some authors (3,4) have contended that medical schools have been negligent in presenting students with the values, skills and attitudes needed to be good decision makers. Caughey has been even more critical than Haney and has suggested that the inadequate treatment of the non-intellectual components of medical education begins with the selection/admission process. The characteristics associated with the physician as decision maker are often ignored by admissions policies (4).

Even though medical educators have tended to ignore the non-intellectual development of students, the process of socialization does occur within medical colleges. Researchers who have examined medical education as a socialization process have identified developmental experiences which have typically resulted in value changes and/or attitude changes (1, 5-8). Some important developmental experiences which have been identified included the dissection of cadavers (1), the death and autopsy of patients (1), the inability to learn all the course

---

Requests for copies should be directed to Dr. Daniels, Department of Guidance & Educational Psychology, 223 Wham, Southern Illinois University, Carbondale, Illinois 62901.

content (9), and the search for a feeling of detached concern for patients (1,7,8). Each of these examples represents an event which is a part of every medical school; therefore, each student has to complete each developmental experience.

Since medical students encounter experiences in school which influence their attitudes and values, socialization is occurring, though it is unplanned and nonsystematic. The central question of socialization theory is: What enduring and relevant changes in moral judgment development are experienced by medical students during their enrollment in medical school?

Two studies have investigated the impact of a medical education on the moral judgment development of medical students. Blizek and Finkler (10) conducted their study in conjunction with an experimental course designed to teach human values in medicine. The course was designed for students in premedical studies, but the experimental design utilized a number of comparison groups, including medical students enrolled in a course in medical ethics. The authors found no changes in the moral development of either the experimental or any of the control groups. In another study, Daniels (11) investigated the moral development of freshman medical students. His study differed from Blizek and Finkler's in that he differentiated among medical students based on the students' preferred interpersonal relationship style (12). Results indicated a difference in the development of moral judgment among students with different patterns of interpersonal relating. Thus, it appears that some students do exhibit significantly modified patterns of moral judgment as a result of operating within a medical school environment, while others do not.

The central question of both studies was: Do medical students adopt new strategies for coping with moral issues as a result of functioning in a highly structured professional environment? Insofar as participation in a program of medical studies involves the socialization of students into a select professional role, it would seem that the transformation of cognitive constructions about moral issues would be an appropriate objective of medical education. If such transformations are not a part of the socialization process, then attempts to incorporate courses concerning medical ethics, medical humanities, or morality and medicine would be useless. If, on the other hand, such transformations are a necessary condition for socialization to proceed, it may be possible to construct curricular materials that foster such cognitive transformations. At present, there is no evidence to support or refute either proposition.

Research Design

The study required the investigation of the longitudinal change of the developmental variable, moral judgment. Data were gathered from the subjects at three separate times over an eighteen month period: (1) Prior to the beginning of classes in the first semester of the first year, (2) at the beginning of the second semester of the first year, and (3) at the beginning of the second semester of the second year. These specific times were identified as critical transition periods in a medical education.

Sample

The sample for this study was comprised of members of the freshman class of the College of Medicine at the University of Iowa for the 1977-78 academic year. A total of 60 students (41 males, 19 females, mean age = 22.5) participated in all phases of the investigation. Participation in all phases was voluntary. Subjects were free to withdraw at any time without penalty.

Instrumentation

Subjects were asked to respond to two standardized instruments. One instrument assessed level of moral judgment development; the other assessed interpersonal relationship style. The Defining Issues Test (D.I.T.) was used to assess moral judgment development. The D.I.T. was originally designed to provide an objective method of assessing Kohlberg's (14) theory of moral development. The D.I.T. is composed of six story dilemmas. After reading each story, subjects are asked to complete three steps. First, they are asked what action the main character of the dilemmas should take. Second, they are asked to rate the importance of twelve issues relevant to the dilemma on a five point Likert scale ranging from great importance to no importance. Finally, they are asked to rank order the four most important issues. Because each issue statement is keyed to a specific stage (type) of moral judgment, subjects receive a score for the following stages: stage 2, 3, 4, 5A, 5B, and 6. Each stage is reported as a separate subscale.

In addition to the six stage subscales, subjects also receive a score for each of two quantitative indices, the P-index and the D-index. The P-index represents the subject's level of principled thinking (i.e., stages 5 and 6). The D-index incorporates information from the subjects' responses to all the issue statements. This characteristic makes the D-index a more comprehensive measure of moral development over all stages (15).

Interpersonal relationship style was assessed by the Fundamental Interpersonal Relationship Orientation-Behavior (FIRO-B), developed by Schutz (12). The FIRO-B measures a subject's interpersonal relationship style by asking the subject to report the typical mode of behavior for each of three interpersonal behaviors: inclusion, control and affection. Each of the three variables is assessed in two ways: (1) Subject's preference for expressing each behavior toward others; (2) Subject's preference for wanting others to exhibit each behavior toward self.

Analysis

A Type 1 analysis of variance (split-plot)(16) was used for the major data analyses. The use of the Type 1 design allowed for the investigation of differences between groups of subjects as well as differences within groups. It was also possible to test for longitudinal changes. A separate analysis of variance was computed for each stage score and both index scores to determine if there were differences in usage across time.

Levels of interpersonal behavior were determined as follows: Each subject's raw scores for each of the scales of the FIRO-B were transformed to z-scores. For each of the scales, subjects were assigned to one of three levels based on their z-score for that scale. If the z-score was greater than +1, the subject was assigned to the High level; if the z-score was less than -1, the subject was assigned to the Low level; all other subjects were assigned to the Average level.

Results

The results of the investigation are reported in two parts. The first part will address the issue of changes in moral judgment development of the sample subjects. The second part will address differences in moral judgment development among subjects based on their preferred style of interpersonal involvement.

Part I. The summary data identifying the average use of each stage of moral reasoning at each of three points on the longitudinal continuum is presented in Table 1:

Table 1
Mean Stage and Index Scores for Time 1, Time 2, and Time 3

Stage	Time 1	Time 2	Time 3	Pattern of Differences
2	2.77	2.40	2.12	---
3	7.32	7.55	6.05	2 > 3
4	14.95	13.80	13.95	---
5A	19.87	19.18	20.28	---
5B	6.77	7.27	8.02	1, 2 < 3
6	5.25	6.22	5.97	---
P	52.59	54.44	56.84	1 < 3
D	28.79	28.87	28.65	---

Significant differences between replications were found for Stage 3, Stage 5B, and the P-index, $p < .05$. Significant differences between replications were not found for any of the other scales. In the follow-up analysis, all pairs means were compared using Tukey's studentized range, $p = .05$. Table 1 indicates which means were significantly different from one another and in which direction they differed. With each scale, the direction of the significant difference was consistent with theory. That is, there was a significant decrease in the usage of less sophisticated (Stage 3) responses and a significant increase in the usage of a more sophisticated and complex (Stage 5B) response. There was also a significant increase in the use of principled thinking, an increase which was due to the increased popularity of responses for Stages 5A, 5B and 6.

Part II. Of the six interpersonal relationship styles, significant results were found for the two affection variables. Due to space constraints, only those two analyses utilizing the D-index will be presented here. The means and standard deviations for three levels of Expressed Affection (EA) for Time 1, Time 2, and Time 3 are reported in Table 2. The results of the analysis indicated a significant main effect for levels of EA ($F = 5.57$; $df = 2.57$; $p < .01$). Neither the main effect for Time, nor the interaction effect for Time by Levels were significant, although the F ratio for the interaction ($F = 2.21$; $df = 4,114$; $p < .07$) did approach significance.

Table 2
Summary of the Data for the D-index: Time 1, Time 2
and Time 3 for Three Levels of Expressed Affection

Levels	N	Time 1 $\bar{X}$	Time 1 SD	Time 2 $\bar{X}$	Time 2 SD	Time 3 $\bar{X}$	Time 3 SD	Total $\bar{X}$
High	11	26.82	5.92	29.27	5.76	27.55	6.26	27.88
Average	37	30.15	5.20	29.38	4.49	30.46	4.68	30.00
Low	12	26.44	4.22	26.92	3.30	24.06	3.96	25.81

A one-tailed follow-up analysis of the significant main effects for levels of EA indicated that there were differences between subjects in the Average and Low levels, $t(57) = 1.86$, $p < .05$.

The test for simple effects for Time 1 indicated significant differences among levels of EA ($F = 3.33$; $df = 2.57$; $p < .05$). The analyses of differences among means indicated differences between the Average level and the Low level, $t(57) = 4.48$, $p < .001$; and differences between the High and the Low levels, $t(57) = 2.06$; $p < .05$.

The test for simple effects for Time 2 was not significant. Significant differences were found, however, among levels of EA for Time 3 ($F = 8.14$; $df = 2.57$; $p < .001$). The analysis of differences among means indicated a difference between the Average level and the Low level, $t(57) = 3.95$; $p < .001$. No other differences among means for Time 3 were significant.

The means and standard deviations for three levels of Wanted Affection (WA) for Time 1, Time 2, and Time 3 are reported in Table 3. The results of the analysis indicated a significant main effect for levels of WA ($F = 4.16$; $df = 2.57$; $p < .05$). Neither the main effect for Time or the Time by Level interaction effect were significant.

Table 3
Summary of the Data for the D-index: Time 1, Time 2 and Time 3 for Three Levels of Wanted Affection

		Time 1		Time 2		Time 3		Total
Levels	N	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$
High	18	28.15	6.32	29.43	6.07	28.31	6.16	28.63
Average	29	29.95	4.36	29.54	3.77	30.56	4.24	30.02
Low	13	27.13	5.30	26.60	3.21	24.82	4.96	26.18

Differences between the Average and Low levels of WA were detected using a one-tailed follow-up analysis. Tests for simple effects indicated significant differences for Time 3 ($F = 5.90$; $df = 2.57$; $p < .005$). The pattern of differences among means was similar to the pattern observed with Expressed Affection. That is, there was a difference between the average and the low levels of WA, $t(57) = 3.30$, $p < .01$. No other differences among means were significant.

Conclusions and Implications

Two conclusions may be drawn from this study. First, the results indicate that on the average the moral reasoning of medical students continues to develop during the first three semesters of medical school. The students demonstrated a greater preference for more complex moral judgments after three semesters of medical school than they did at the time of their matriculation. Second, the results demonstrate that the manner in which students adapt to their social environment influences their moral development. Subjects grouped in the Average level of both Expressed and Wanted Affection demonstrated higher scores on the moral judgment index.

The findings of both Affection variables were as expected. The Affection variable is concerned with the need for involvement in nonsuperficial relationships. People with average affection scores, both wanted and expressed, are comfortable in close, caring relationships, but do not need such relationships to be fulfilled. In contrast, people with low affection scores are uncomfortable with close, intimate relationships and are reluctant to be involved in them (17). Morality is concerned with guiding the conduct of interpersonal relationships.

Thus, it may be inferred that people who are comfortable with sensitive interpersonal involvements demonstrate greater development on a variable that is rooted in social relationships, i.e., morality.

These results have important implications for medical educators. Two will be mentioned here. First, because of the evidence of continuing moral development during medical school, interventions designed to promote such development may be established and evaluated. Second, because interpersonal relationship style appears to influence the development of moral judgment, it seems important that programs of moral education should consider differences in the social adaptivity patterns of students. More research is needed to verify these conclusions.

References

1. Levinson, O.J. Medical Education and the Theory of Adult Socialization. J. Health Soc. Behav., 8:253-265, 1967.
2. Caughey, J.L. Nonintellectual Components of Medical Education. J. Med. Educ., 48:619-625, 1967.
3. Haney, C.A. Psychosocial Factors Involved in Medical Decision Making. In Coombs and Vincent (Eds.), Psychosocial Aspects of Medical Training. Springfield, Illinois: Charles C. Thomas, 1971.
4. Ferguson, M. Medical School Reform. Human Behavior, 7:43-45, 1977.
5. Becker, H.S., Geer, B., Hughes, E.C., Strauss, A.L. Boys in White. Chicago: The University of Chicago Press, 1961.
6. Coombs, R.H., and Boyle, B.P. The Transition to Medical School: Expectations versus Realities. In Coombs and Vincent (Eds.), Psychosocial Aspects of Medical Training. Springfield, Illinois: Charles C. Thomas, 1971.
7. Fox, R.C. Training for Uncertainty. In Merton, Reader, and Kendell (Eds.), The Student Physician. Cambridge: Harvard University Press, 1957.
8. Fox, R.C. The Medical Students' Training for Detached Concern. In Lief, Lief, and Lief (Eds.), The Psychological Basis of Medical Practice. New York: Harper and Row, 1963.
9. Coburn, O. and Jovaisas, A.V. Perceived Sources of Stress Among First Year Medical Students. J. Med. Educ., 50:589-595, 1975.
10. Blizek, W.L. and Finkler, D. Teaching Human Values in Medicine. J. Med. Educ., 52:858-859.
11. Daniels, M.H. Moral Judgment Development as a Function of Interpersonal Relationship Style: An Exploratory Study. Unpublished doctoral dissertation, University of Iowa, 1978.
12. Schutz, W.C. The Interpersonal Underworld. Palo Alto: Science and Behavior Books, Inc., 1966.
13. Rest, J.R. Manual for the Defining Issues Test. Minneapolis: University of Minnesota, 1974.
14. Kohlberg, L. Stage and Sequence: The Cognitive-Developmental Approach to Socialization. In Goslin (Ed.), Handbook of Socialization Theory and Research. Chicago: Rand McNally, 1969.
15. Davison, M.L. and Robbins, S. Indexing Moral Development. Unpublished manuscript, University of Minnesota, 1977.
16. Lindquist, E.F. Design and Analysis of Experiments in Psychology and Education. Boston: Houghton-Mifflin, 1953.
17. Ryan, L.R. Clinical Interpretation of the FIRO-B. Palo Alto: Consulting Psychologist Press, Inc., 1977.

MORAL REASONING AND PHYSICIANS' DECSIONS IN CASES OF CRITICAL ILLNESS

Daniel Candee, PhD, T. Joseph Sheehan, PhD, Charles D. Cook, MD, Susan R. Husted, BS

Research in Health Education, University of Connecticut Health Center[1]

Among the most important decisions a physician makes is the type of treatment he or she will employ in cases of critical illness. Such decisions lie at the intersection of both medical and moral concerns. Questions of whether to provide heroic care, care that is often expensive, painful, and unfortunately often futile, are not simply matters of technical expertise. They involve value choices.

How does a physician make decisions on such matters? What factors are most influential in his or her thinking? Answers to some of these questions were provided by Crane (1975) who surveyed 922 practicing physicians and residents in pediatrics and in medicine, concerning their attitudes towards treatment of the critically ill. Crane found that the degree of activism reportedly engaged in by such physicians differed greatly depending on such case factors as the possibility for recovery, the type of damage sustained, and the attitude of the patient or family towards aggressive medical treatment and on such physician factors as professional status, prestige of hospital affiliation and religious background, While such findings clarify the considerations likely to be involved in a physician's decisions they do not tell us whether these considerations are part of any larger psychological structures.

The possibility that such elements stem from an individual's general structures of moral reasoning is suggested by a growing tradition of research in that field. Work by Kohlberg (1969, 1975) and his associates has shown that the underlying principles or "cognitive structures" which an individual uses to resolve moral dilemmas develops through a series of six sequentially ordered stages. The stages begin with a punishment and obedience orientation (stage 1) and develop through a conventional acceptance of societal or in-group rules (stages 3 and 4), to a post-conventional autonomy in which moral decisions are made on the basis of universally valid ethical principles (stages 5 and 6). At each succeeding stage there is a greater appreciation of the welfare of others and a greater desire to resolve moral dilemmas in a fair and equitable manner.

When applied to the realm of decision-making in cases of critical illness the implication of this research is that higher stage reasoning, since it is based on a greater respect for personal dignity and individual rights, will lead physicians holding such reasoning to be more sensitive to characteristics of the patient and of the disease which affect those rights. In order to test this proposition we conducted a survey of 452 residents and practitioners in which we measured the relationship between moral reasoning and attitudes towards the treatment of critical illness. We hypothesized that physicians of higher stage moral reasoning would be more willing to limit their use of active treatment in situations where the patient (or patient's family) requested such a limit and/or where the quality of life likely to be led after therapy was so low as to violate the patient's dignity as a human being.

---

[1] First author's address: Center for Moral Education, Larsen Hall, Harvard University, Cambridge, Massachusetts 02138.

Method

Subjects. Questionnaires measuring both moral reasoning and decisions in cases of critical illness were distributed to 693 physicians, in pediatrics and internal medicine. Since all but 74 of the recipients were in pediatrics only the pediatricians will be reported here. A total of 452 subjects (76 percent of the pediatric recipients) returned completed questionnaires. Two hundred fifty-six of the subjects included in the final sample were currently in residency while 186 were practicing physicians. One hundred and one of the residents were graduates of American medical colleges while 155 came from foreign schools. All practitioners were U.S. medical school graduates. Since the samples were neither random nor stratified results must be interpreted with reference to those subjects actually studied.

Measures. The first questionnaire, devised by Crane (1975), assessed the physician's reported degree of activism in the treatment of critical illness based on a series of six simulated cases involving infants born with severe defects. The questionnaire systematically varied six factors: prognosis (salvagable vs. unsalvagable), type of damage (mental vs. physical), family attitude (high, i.e., positive towards active treatment vs. low, i.e., negative towards active treatment), socio-economic status (SES) of the patient's family, type of error (doctor's error vs. machine error), and the opportunity for research (yes vs. no).

In order to provide counterbalanced conditions for every story without presenting any individual subject with both conditions for a given story two forms of the questionnaire were used. For example, the first story, A2, a case of an infant with high lumbar myeolmeningocele, appeared on Form A of the questionnaire only. The case is one of physical damage, salvagable prognosis, and high family socio-economic status (parents described as "well-educated and financially comfortable"). The family's attitude towards active treatment was not specified. The same case also appears on Form B although here the family is described as "not having finished high school" (low SES). Any one subject received either Form A or Form B. Following each case there appeared a series of seven possible choices for action, ranging from minimal support (e.g., intravenous) to heroic treatment (e.g., resuscitation). Respondents were asked to indicate whether they would, maybe would, or would not perform each type of action.

A subject's activism score on each story was determined by awarding 3 points for each action the respondent indicated he or she would perform, 2 points for each action marked "maybe," and 1 point for each action that would not be performed. An individual's story activism score was able to range from 7 to 21. An index of average total activism was computed across the six stories. A third activism index reflecting intra-individual variability was computed by determining the standard deviation for each subject across all stories completed.

The second instrument, the Defining Issues Test (DIT) is a structured measure designed by Rest (1976) on the basis of Kohlberg's theory of cognitive moral development. Subjects are presented with six moral dilemmas and are asked, for each dilemma, to choose from a set of twelve considerations those which the subject believes are most important in resolving the issue at hand. The considerations have been pre-selected to represent several of Kohlberg's moral stages. An individual's score is expressed in terms of the percentage of responses which have been keyed at the principled stages of moral reasoning (stages 5 and 6) to which the subject gives high priority.

Results

Before reporting the data which bears on the relationship of physician attitude to moral reasoning let us examine the absolute level of activism found in our sample. Mean activism scores for the 12 pediatric questions are presented in Table 1. Since the average activism score across all stories was significantly different for residents than for practitioners (see column 1, Table 1) $p<.05$, we report the results for each group separately as well as combined. As the combined mean row indicates the highest degree of activism was found on story A5, insufficient clamping of the umbilical cord (19.70), and the least found on story A1a, anencephaly, (9.74). Not only were these stories the most extreme for the combined sample but they were also most extreme for the separate samples as well (see residents means and practitioners means).

Since the stories on the Crane questionnaire follow a design structure they can be combined to reveal larger effects. For example, a variable that measures activism when family attitude is positive can be constructed by averaging the means of story A4 and B3. This variable can be contrasted with the mean activism score obtained when family attitude is negative (stories A3 and B4).

Differences between the following contrasts were found to be significant ($p<.01$) for the individual as well as the combined groups: 1) family attitude (pro-activism being treated more actively than anti-activism), 2) researchability (the presence of such opportunity being treated more actively than the absence of such opportunity), and 3) type of error (doctor's error being treated more actively than mechanical error). No significant differences were found for conditions of high and low family SES. A confirmation sample of 81 residents, using a slightly enlarged version of the Crane questionnaire confirmed the family attitude and researchability contrasts but did not confirm the type of error contrast.

We turn now to the relationship between moral reasoning and physicians' attitudes towards the treatment of the critically ill. Correlations between the DIT and activism on pediatric stories appear in Table 2. As can be seen the degree of activism averaged across all stories in Form A and across all stories in Form B is negatively related to moral reasoning among residents ($r=-.41$ for Form A and $-.23$ for Form B) and either unrelated or positively related to moral reasoning among practitioners ($r=-.05$ Form A, $.27$ Form B). Due to the difference in average activism correlations for residents and for practitioners we report the results on individual stories separately for the two groups.

An examination of individual story correlations in Table 2 reveals that for residents, the lesser degree of activism advocated by the more principled subjects is due almost entirely to those cases where either salvagability or family attitude is low (i.e., stories A1, B1, A3, and B4). Thus, our original hypothesis concerning the direction and nature of the moral reasoning-activism correlation is supported, at least among residents. As predicted, the difference between high and low moral reasoners is most apparent when circumstances such as the parents' wishes or the patient's inevitable death become reasons for some physicians to depart from their usual practice of maintaining life.

Summary verification of our major hypothesis, among residents, is indicated by the significant correlation between moral reasoning and a measure of the variability in response to the six pediatric stories (see average variability column). The greater variability among high stage reasoners can be interpreted

TABLE 1

MEAN ACTIVISM SCORES FOR PEDIATRIC STORIES

Prognosis		Salvagable								Unsalvagable			
Type of damage		Physical		Mental						Physical		Mental	
Family attitude [a]				Fam +		Fam -							
Family SES or Type of error or Researchability		S+	S-	S+	S-	S+	S-	D Error	M Error	R+	R-	R+	R-
Story #	Avg. act.	A2	B2	A4	B3	A3	B4	A5	B5	B1	B1	A1	A1
Residents													
Mean	15.41	15.79	14.38	16.67	18.20	15.19	14.50	19.51	18.89	16.95	15.63	15.01	10.71
SD	3.42	3.29	4.24	3.15	3.03	4.16	4.50	1.63	2.35	3.04	3.85	4.28	3.61
Practitioners													
Mean	14.51	15.68	14.56	15.64	17.10	12.41	12.41	19.84	18.56	16.40	14.03	14.51	8.79
SD	3.44	3.96	4.07	3.91	3.00	4.17	4.44	1.79	1.93	3.21	3.78	4.50	2.55
Combined													
Mean	15.05	15.73	14.95	16.10	17.50	13.80	13.50	19.70	18.70	16.60	14.83	14.75	9.74
SD	3.43	3.62	4.15	3.53	3.02	4.17	4.47	1.71	3.10	3.12	3.82	4.39	3.08

Note: N= Residents (85 Form A/ 100 Form B), Practitioners (76 Form A/78 Form B), Combined (161/178).

The content of the stories is as follows (story numbers precede description):
1 - Anencephaly (A1/Hypoplastic ventricle (B1).
2 - High lumber myelomeningocele with no nerve function in legs, bladder, or rectum.
3 - Downs syndrome with severe respiratory distress.
4 - Premature separation of the placenta resulting in marked spasticity and hypertonia.
5 - (A5) insufficient clamping of umbilical cord stump by doctor with resulting bleeding/ (B5) 1400 gram infant is exposed to temperature of 107 degree Fahrenheit for at least 30 minutes due to mechanical failure of incubator with febrility, marked lethargy, lack of reflexes, dehydration, and acidotia resulting.

[a] Where blank family attitude is unspecified.

TABLE 2

CORRELATIONS BETWEEN MORAL REASONING AND ACTIVISM ON PEDIATRIC STORIES

	Average Activism		Average Variance		Stories[a]											
	Fm A	Fm B	Fm A	Fm B	A2	B2	A4	B3	A3	B4	A5	B5	B1a	B1b	A1a	A1b
Residents	-.41*	-.23*	.39*	.26*	-.14	-.35*	-.25*	.11	-.27*	-.40*	.05	.07	-.24*	-.30*	-.34*	-.40*
Practitioners	-.05	-.27*	-.02	-.08	-.06	.23*	-.03	.12	-.19*	.11	-.01	.25	.18	.18	-.03	.01

Note: N = Residents (62 Form A, 59 Form B), Practitioners (85 Form A, 82 Form B). Fewer Subjects are reported in Table 2 than in Table 1 due to incomplete DIT's.

[a] For structure of stories see Table 1.

* $p<.05$

as demonstrating that such residents were more sensitive to characteristics of each situation and varied their treatment accordingly. In comparison, lower stage residents seem to treat cases more uniformly.

Before leaving the topic of the moral reasoning-activism correlation we must consider a competing hypothesis. Perhaps the relationship between the variables can be explained by differences in institutional policies. In order to test this proposition a two-way analysis of variance was conducted using moral reasoning (high, medium, and low) and type of institution (university-prestige, university, and community)as factors and average activism as the dependent variable. Results of this analysis within Form A revealed two significant main effects and no interaction. Within Form B only a significant type effect was found. Thus, it appears that some but not all of the observed relationship between moral reasoning and activism may either be due to, or be mediated by, differences in institutions.

Discussion and conclusions

Two major conclusions seem warranted on the basis of our findings. One is that pediatricians' decisions concerning the treatment of critical illness are at least partially associated with broader structures of ethical thought. Among residents, those subjects who were shown to reason in terms of universal ethical principles were most likely in particular cases to tailor their treatment of critical illness to both the explicit and implicit rights of patients. This was shown by correlations between DIT scores and the family attitude factor (explicit rights) and the salvagability factor (implicit rights). The second conclusion is that pediatricians as a whole are most influenced by family attitude in deciding to treat salvagable cases and by the possibility of research in deciding to treat unsalvagable ones.

The implications of our results are primarily educational. First, since specific moral decision-making has been shown to be associated with broader moral structures it would seem that programs in medical ethics should encourage physicians to examine the broader ramifications of their basic moral concepts (for example, the justification of human rights, the nature of human dignity and the basis of moral principles). Ethical discussions should not be limited to solving cases alone. Second, our findings suggest that since family attitude is the most influential factor in the decision to treat critical illness, physicians should receive special instruction on how to help families both recognize their true preferences and communicate them to the physician.

References

Crane, D. The Sanctity of Social Life: Physicians' Treatment of Critically Ill Patients, New York: Russell Sage Foundation, 1975.

Kohlberg, L. Stage and sequence: the cognitive developmental approach to socialization, in D. Goslin (ed.) Handbook of Socialization Theory and Research, Chicago: Rand-McNally, 1969, 347-480.

Kohlberg, L. The cognitive-developmental approach to moral education. Phi Delta Kappan, June, 1975.

Rest, J. Moral Judgment Related to Sample Characteristics: The Defining Issues Test Related to Age, Education, IQ, Geographical Region, Religious Affiliation Sex, and Socioeconomic Status. Final Report to the National Institutes of Mental Health, Grant # 8703 MH 24988, 1976.

COPING AND EGO DEVELOPMENT IN INTERNSHIP

Julie C. Donnelly, Ed.D., Instructor, Department of Research in Health Education*
University of Connecticut School of Medicine
Farmington, Connecticut 06032

Statement of the Problem

Through an in-depth inquiry into personal and professional development in young physicians, the intent of this study was to provide a descriptive account of the internship experience and to examine the degree to which individuals at differing developmental stages (as measured by Loevinger's stages of ego development) experience the same life event (internship) differently. The questions with which I began were the following: Do ego level differences constitute sufficiently distinct orientations such that stage-related patterns could be found within structured interview material and self-reported coping mechanisms? Do ego level and/or coping style bear any relationship to both self-assessments of the experience and faculty ratings of clinical performance?

Background

Loevinger's stages of ego development (1966, 1973, 1976) describe a progression of characterological types and a development sequence for tracing changes in an individual's frame of reference for interpreting and responding to daily experience. Loevinger (1976) makes the claim that ego development has two components--meaning construction and mastery. However, her model does not assume to predict any relationship between ego development and overt behavior; Loevinger herself explicitly disavows the likelihood of finding specific behavioral signs at a given level (Loevinger and Wessler, 1970). Both Hauser (1976) and Lasker (1978) suggest that Loevinger's descriptions of character styles unique to each stage imply the existence of stage related behavioral patterns. In addition, due to the paucity of work on the application of her model in educational and clinical settings, I anticipated that relating ego level to behavioral differences in functioning within medical training would represent a useful practical addition.

Loevinger's stages can be described briefly as follows:

Impulsive--An action is bad because it is punished
Self-Protective/Opportunistic--Events and people are seen in terms of exploitation and expedience; life is a zero-sum game
Conformist--Rules are partially internalized; mutual trust only between in-groups; preoccupation with appearances
Self-Aware--Beginning of self-reflection and recognition of alternative possibilities
Conscientious--Evaluates rules and allegiances; internalized morality; capacity for self-criticism; preoccupied with obligations
Individualistic--Increased concern with individuality, emotional dependence, and complexity of experience
Autonomous--Appreciation of paradox and ambiguity; copes with inner conflict; recognizes others' autonomy
Integrated--Reconciles conflicting demands; renounces the unattainable; cherishes individual differences

---

*Address reprint requests to Dr. Donnelly

Within the extensive literature on coping, defined here as the ways individuals respond to stress, there is a staggering diversity of approaches and opinions. A taxonomy of coping developed by Lazarus (1966, 1975) emphasizes two major functional domains, Non-Palliative and Palliative Modes. Non-Palliative modes include coping by direct problem solving actions which respond to the stress situation itself, for example, behaviors relevant to defining, attacking, or meeting the task at hand. Palliative modes of coping refer to thoughts or actions whose goal is to relieve the stress response or the emotional impact of stress (i.e., bodily or psychological disturbances). The term palliative implies that these methods do not actually alter the threatening or damaging events but they make the person feel better. According to this classification scheme, most people employ complex combinations of both non-palliative and palliative methods to cope with stress.

If Loevinger's scheme is divided into Level I (Stages Impulsive to Self-Aware) and Level II (Stages Conscientious to Integrated) I hypothesized that different styles of coping were likely to be associated with these two levels. To the degree that individuals at Level I are far more preoccupied with externalizing and externalities I expected them to regard stress as primarily externally induced and their primary coping mechanisms to be action oriented and directed at the environment rather than themselves. In contrast, to the degree that individuals at Level II are more oriented to internal rules and standards and internalizing responsibility, then I expected this group to regard stress as a by-product of both the external situation and themselves, and to therefore cope by trying to alter both the environment and the self, with much emphasis on their own feeling stage. Returning to the distinction between Non-Palliative and Palliative Modes, I therefore hypothesized that the Level I group would exhibit a tendency to use primarily Non-Palliative Modes, while the Level II group would be more likely to use a combination of both Non-Palliative and Palliative Modes with an emphasis on the latter. Staged another way, my hypothesis was that the Level I group would be less likely than the Level II group to employ palliative modes of coping.

Methodology

Twenty young, male physicians comprised the sample for this study. Each was completing his internship in a training program affiliated with the University of Connecticut School of Medicine in 1976-77. Subjects were assigned to either ego Level I or II on the basis of their scores on Loevinger's 36 item Sentence Completion Test. The assumption of the scoring paradigm is that each person has a core level of functioning, and that each sentence item is an independent eliciting event for revealing that core level. The scoring methodology is described by Loevinger and Wessler (1970). Interrater correlations as reported by Hauser (1976), Hoppe (1972), and Cox (1974) fall within the range of .76 and .86. Waldman (1975) probed the reliability of the SCT itself and reported test-retest correlation of total protocol stage scores varying from .44 to .79. Item sum scores are reported of .64 and .91. Split half reliability for the two groups was .90 and .85, while internal consistency coefficients ranged between .80 and .89. Hauser (1976) reviewed the research on the discriminant validity and construct validity and found a variety of evidence for both types.

All subjects were interviewed extensively about their internship experience. All interviews were transcribed and the material was analyzed for differences between the two ego level groups by examining the responses to clusterings of questions on related topics. In addition to an overall protocol analysis, I also extracted from each interview all examples of self-reported observations of strategies for coping with the internship experience. Each of these were in turn sorted into the two primary coping domains of Non-Palliative and Palliative modes, and then further into subcategories of each. A second rater coded and categorized every fifth page of the interview material plus all questions concerned directly with coping. Interrater agreement for locating the individual coping statements within the interviews was 75%. The interrater agreement for the classification between Palliative and Non-Palliative was 92%. Within the Palliative and Non-Palliative domains agreement was 100%.

In addition, a measure of each subject's clinical performance was employed in order to provide a composite portrait of each subject on 18 aspects of clinical behavior judged to be most important by a group of university faculty and house officers. The particular rating form was based on earlier work by Margolis and Cook (1974), and modified by Sheehan (1978b, 1979). Each intern was ranked on characteristics which were categorized under one of three competency areas: PERSONAL CHARACTERISTICS such as honesty, dependability, acceptance of responsibility, INTERPERSONAL SKILLS such as ability to relate to patients and work with others, and PROFESSIONAL ABILITIES AND KNOWLEDGE such as knowledge of medicine, awareness of own limits, and level of clinical judgement. Each intern was rated by three to eight faculty members. Margolis and Cook report reliabilities in the .75 range. Sheehan (1979) reports an average intercorrelation between raters of .67. Clinical performance was certainly not the only way in which individual differences in coping styles could have been assessed from the point of view of adequacy of functioning; it was chosen because it represented an accessible performance criterion based on actual behavior, as judged by raters other than the author. It was also of interest because of the growing concern in identifying meaningful, non-cognitive predictors of clinical performance.

The statistical technique of discriminant analysis was employed in order to investigate the relationship between level of ego development and coping repertoire. A variety of both parametric and non-parametric techniques were utilized to relate the results of this analysis to the clinical performance ratings.

Results

Variations in the Internship Experience

Following a delineation of the major areas of reported stress, the interns' self-assessments of their functioning at work and home were analyzed. The primary difference between the two ego level groups was an overall tendency for those interns in Group II to exhibit a greater awareness of complexity and of emotional impact, whether in the situational, interpersonal, or psychological realm. As compared to those in Group I, they reported a greater sense of discrepancy between the ideal and the real and were consequently less content with the internship experience as it presently exists.

Coping Style and Ego Development

Through the discriminant analysis a strong relationship was found between membership in Group I and a coping style which relied more heavily on Non-Palliative coping and membership in Group II with a coping style which relied more heavily on both Non-Palliative and Palliative coping. This relationship was so strong that for any individual subject, coding his interview for coping style alone was found to predict membership in ego level I or II with 90% accuracy.

Coping Style, Ego Development and Clinical Performance

When self-reported coping style, level of ego development and clinical performance ratings were combined, several interesting patterns emerged. On initial examination of the relationship between ego development and clinical performance, although there appeared to be a trend toward higher performance scores in Group II this did not reach statistical significance. When the subjects were divided into three ego level groups (high, middle and low) of approximately equal size, this produced a curvilinear pattern whereby a high performance rating was far more likely to be associated with an ego development score which was either extremely high or low than it was with a mid-range score. This relationship was analyzed by a Fisher Exact Test and was found to be significant at the $p = .05$ level.

Since this neither made sense from a theoretical nor intuitive point of view, the data was examined further to discover what further explanation of this pattern could be found. When the relationship between coping style (as indicated by an individual's discriminant score on the Palliative--Non-Palliative coping dimension) and performance rating was examined, this produced a Spearman rank correlation of only $r = 0.12$. A correlation coefficient of $r = .33$ was achieved by ranking all performance scores as either high or low, but such a grouping still did not adequately account for a relationship.

The next step in this analysis was to investigate whether a closer examination of this pattern might account for a clearer relationship between ego level and performance rating than achieved so far, on the assumption that some combination of ego level, discriminant score and coping repertoire might bear a stronger relationship to work performance than any of these when considered alone.

When each of the scores on all subjects were displayed by ranking discriminant score and performance rating, visual observation suggested a trend which can be described as follows:

1. At the Conscientious Stage and above if a subject's coping repertoire included two or more Palliative categories with at least two entries in each of these categories, he was likely to be in the higher performance group. Without these Palliative scores, an individual would more likely be in the lower performance group.
2. At the Self-Aware Stage an individual must have coping scores in three or more categories with frequencies greater than two to be in the high performance group. If not, he was likely to be in the lower performance group.
3. At the Conformist Stage and below an individual was likely to be in the lower performance group regardless of his coping repertoire.

The decision rules which established cutoff points for predicted membership in either the higher or lower performance group differed by whether one scored at a high, middle, or low level of ego development: the criteria became more stringent as ego stage decreased. In essence, the power of ego stage alone in contributing to clinical performance differed at each of three levels. Within each level, a broader coping repertoire, and in particular the presence of Palliative scores, was the determining factor between a higher or lower performance rating.

After generating this model it was necessary to actually test the data to discover how closely it fit the suggested pattern. With 17 out of 20 or 85% of cases classified correctly, this model appeared to offer significant predictive power. A Fisher Exact Test of this data indicated that the null hypothesis of no significant differences between groups could be rejected at the p = .005 level, thus providing strong support that the data fit this model very well. In addition, it suggested that the differential patternings indicated by this data were describing real distinctions between these groups.

Discussion

On a theoretical level the data strongly suggest that a linkage can be made between the way in which an individual makes sense of his experience (meaning-making) and how he behaves and how well he performs (mastery). The way in which individual interns describe themselves making sense of their work experience, and the way in which they describe their own coping strategies appears to be intimately connected to their level of ego development.

The major extension of the theory, however, is that both ego development and coping bear a clear relationship to actual behavior as measured by clinical performance. An individual at the Conscientious level and above is more likely to have a high performance rating than someone at the Self-Aware level and below. A coping style which comprises both Palliative and Non-Palliative modes is also more likely to be associated with high performance than one which relies on only the latter. However, it is the combination of both of these which is the most predictive of a higher performance rating. An ego level of Conscientious or higher, together with a coping style which gives heavy emphasis to one's affective experience, is most predictive of a high performance rating. Conversely, an ego level of Conformist or below is least likely to be associated with a high performance rating. At the ego level mid-range of Self-Aware, a coping repertoire which is developed in both the Palliative and Non-Palliative realms is most likely to lead to a high performance rating.

Implications

Within the current interest in the relationship of noncognitive characteristics to clinical performance, the empirical relationship found here between level of ego development and adequacy of clinical performance should be taken as further evidence of their significance. Because of the strong relationship between Loevinger's stages of ego development and Kohlberg's (1969) stages of moral development, it provides further support for the relationship between moral maturity and clinical performance as reported by Sheehan (1978, 1979) and Husted (1978).

Conclusion

The data presented here strongly support the personal and professional advantages for physicians who possess the psychological sensitivity to affective and interpersonal nuances, as well as the cognitive complexity and the instrumental skill to deal directly with the structural aspects of a situation. Level of ego development, together with the breadth and balance of one's coping repertoire in both the palliative and non-palliative realms, appear to be crucial elements in interns' clinical performance and their self-assessed functioning both at work and at home. The evidence in this study also supports the conclusion of Werner et al., (1979) that those interns who were least able to deal with stress were those who received the poorest performance ratings, and that putting doctors in a situation in which they must suppress their feelings may increase the risk of poor adaptation and poor attitudes. It becomes critical to consider implications for the design of medical education, including direct structural changes in internship, possibilities for enhancing individual development and coping abilities.

123

References

Cox, N. "Prior Help, Ego Development, and Helping Behavior," Child Development, 45, 594-603, 1974.

Hauser, S. "Loevinger's Model and Measure of Ego Development: A Critical Review," Psychological Bulletin, 83, 928-955, 1976.

Hoppe, C. "Ego Development and Conformity Behavior," Doctoral Dissertation, Washington University, 1972.

Husted, S. "Assessment of Moral Reasoning in Pediatric Faculty, House Officers and Medical Students," Research In Medical Education Conference Proceedings, pp. 439-441, 1978.

Kohlberg, L. "Stage and Sequence: The Cognitive Developmental Approach to Socialization," In D. Goslin (Ed.) Handbood of Socialization Theory and Research, Chicago, Rand McNally, 1969, pp. 347-480.

Lasker, H. "Ego Development and Motivation: A Cross-Cultural Cognitive Developmental Analysis of Achievement," Doctoral Dissertation, University of Chicago, 1978.

Lazarus, R. Psychological Stress and the Coping Process, New York, McGraw-Hill, 1966.

"A Cognitively Oriented Psychologist Looks at Biofeedback," American Psychologist, 30, 553-6, 1975.

Lieberman, M. "Adaptive Processes in Late Life," In N. Datan and L. Ginsberg, eds., Life Span Developmental Psychology, New York, Academic Press, 1975, pp. 135-159.

Loevinger, J. "The Meaning and Measurement of Ego Development," American Psychologist, 21, 195-206, 1966.

"Recent Research on Ego Development," Address to Society for Research in Child Development, Philadelphia, March 1973.

Ego Development, San Francisco, Jossey-Bass, 1976.

Loevinger, J. and Wessler, R. Measuring Ego Development I and II, San Francisco, Jossey-Bass, 1970.

Margolis, C.Z. and Cook, C.D., "Rating Pediatric House Officer Performance," Pediatric Research, 8, 472, 1974.

Redmore, C. and Waldman, K. "Reliability of a Sentence Completion Measure of Ego Development," Journal of Personality Assessment, 39, 236-243, 1975.

Sheehan, T.J. "The Relationship Between Moral Reasoning and Clinical Performance," Research In Medical Education Conference Proceedings, pp. 444-5, 1978.

"Moral Judgement as a Predictor of Clinical Performance," A.E.R.A. Presentation, April 1979.

Werner, E., Adler, R., Robinson, R., and Korsch, B. "Attitudes and Interpersonal Skills During Pediatric Internship," Pediatrics, Vol. 63, pp. 491-99, 1979.

PRECIS

PERFORMANCE EVALUATION IN GRADUATE MEDICAL EDUCATION

MODERATOR: Fredric D. Burg, M.D.
National Board of Medical Examiners

UTAH CONTINUING STUDIES ON PHYSICIAN AND PHYSICIAN-IN-TRAINING PERFORMANCE

The feasibility of obtaining multiple measures of physician characteristics and performance was demonstrated first on a study of 167 physicians-in-practice and second on a sample of interns and residents-in-training in Surgical and Pediatrics specialties. The measures were generally found to be of psychometric quality, so implementation efforts can be on target, at all levels of selection and training, toward producing a gerater percent of excellent physicians in practice.

FAILURE OF EXAMINATION ANSWERS TO EVALUATE ACTUAL PRACTICE PATTERNS BY MEDICAL HOUSESTAFF IN AN OUTPATIENT CLINIC

This study fails to find a correlation between physicians' performance on an examination with performance in an outpatient medicine clinic. The examination dealt with the use of screening tests.

RELIABILITY IN FOLLOW-UP EVALUATIONS OF GRADUATES AS RESIDENTS

This paper presents a model for the design of residency follow-up studies aimed at collecting data on the performance of medical school graduates. The findings indicate strong agreement and high reliability of supervisors' ratings.

THE EFFECTS OF CLINICALLY RELEVANT MULTIPLE-CHOICE ITEMS ON THE STATISTICAL DISCRIMINATION OF PHYSICIAN CLINICAL COMPETENCE

This study investigated the effect of clinically relevant multiple-choice item content on the discriminant validity of subscale scores intended for specialty certification. Clinically relevant items were more discriminating of criterion groups, more reliable, and lower in mean item difficulty than were items selected for medium difficulty.

UTAH CONTINUING STUDIES ON PHYSICIAN AND PHYSICIAN-IN-TRAINING PERFORMANCE*

by

Dominic Albo, M.D., Calvin W. Taylor, Ph.D.,
Philip Price, M.D., and Brent Page, M.S.

College of Medicine and Department of Psychology
University of Utah

I. PURPOSE

Our studies on the Measurement of Physician Performance of those practicing in Utah have initially emphasized basic research and development work on the characteristics of excellent physicians. More recently, as we have continued into a third decade, the basic research and development work have continued, together with a strong, sustained, five-year effort in applying and implementing these findings by evaluating and providing feedback on the performance of surgical and pediatrics interns and residents in training. The broad purposes have been to determine what is an excellent physician in practice and to develop sound measuring instruments of such physicians' performances and characteristics; then to retool these measures to make them applicable at the specialty training level -- and eventually at the earliest stages of premedical and medical school selection and education. The ultimate intent is to increase the number of physicians who perform at an excellent level in their professional practices and careers. Our conviction is that this gets at the root of the problem, namely, there is nothing more basic to the entire field of Medicine than physicians and how well they perform in their various roles and activities.

II. CRITICAL REVIEW OF THE LITERATURE

Because the problem we are tackling is very broad, we have done lengthy reviews of the literature in our two large earlier research reports. One review is in chapter one, pages 3-29, of our first major study titled Measurement of Predictors of Physician Performance (Price, et al., 1971). The second review is in chapter two, pages 7-18, of Comprehensive Analysis of Physician and Physician-In-Training Performance, volume I (Taylor, Nelson, and Price, 1974). Relevant to physicians-in-training is the article entitled "Multifactor Evaluations of Surgical Trainees and Teaching Services" (Albo, et al., 1976). A very recent article pertinent to the entire matter has been written by Harrison G. Gough, who has also sustained a considerable effort for many years in this general area. He reviews the literature and reports his best recommendations in his article on "How to Select Medical Students" (1979).

---

*For reprint requests, write to Calvin W. Taylor, Professor of Psychology, Building 404, University of Utah, Salt Lake City, UT 84112.

III. NATURE OF THE METHODOLOGY AND STUDIES

A. Physicians in Practice. Our latest and largest study on the measurement of physician performances will be briefly reported. In a three-year effort with Public Health Service support (Contract HSM 110-71-171), we intensively applied to physicians in practice the 87 positive and 29 negative (116) characteristics found in our large earlier study cited above. We obtained over 400 items of information on these characteristics on a total of 167 physicians, namely, 34 in Surgery, 35 in Pediatrics, 34 in Internal Medicine, 31 Rural GPs, and 33 Urban GPs.

Multiple sources of measurement were used, namely (1) physician observers who spent a day observing every physician studied, together with examining his patient records and watching him in the hospital and office; (2) observer-ratings from multiple nurses per physician; (3) observer-ratings from one or more residents and interns (whenever available); (4) observer-ratings from multiple patients; and (5) self-ratings from the physician.

A subsample of the 116 characteristics was measured by each of the sources in terms of the characteristics of performances that they could honestly observe and soundly measure. This meant that certain items of information were measured by all of the sources, others by some of the sources, and some measured by only one of the sources. A separate measurement form was constructed to be pertinent to each of these five sources. In addition, various kinds of biographical, demographic, and objective performance data were collected by the five forms. Scores developed across the items were evaluated to determine their psychometric qualities in terms of reliability and validity of criterion measures.

B. Interns and Residents in Training. We have been studying 50 surgical trainees a year. Over their five years of training, the number decreases considerably from about 20-25 first-year trainees (interns) up to only a handful or so of fifth-year trainees (Chief Residents). In pediatrics we are studying about 25 trainees, nearly all of whom eventually finish a three-year residency training in that specialty.

New relevant measures have been used with feedback. A main intent is to see that the trainees become more physician-like than student-like as they get their post-doctoral specialty training. Considerable effort has been spent in developing and revising measures of the characteristics and performances of interns and residents. Revisions have been necessary to satisfy all the needed purposes and conditions. Trainees have helped determine that the forms measure the attributes and performances that are relevant to their training. The faculty and practicing physicians in the ten different training situations in multiple hospitals have to be willing and able to fill out the measurement forms which also need to be designed so they can be readily processed to yield feedback information to the trainees, for sound department records, for continuing and future research, and for use in personnel decisions regarding retention each succeeding year. Confidentiality was designed into the forms and the data processing system. Feedback scores were eventually generated every six months, thus covering two-month periods in three different hospital training situations. The data have been optic scanned and computer scored and processed in various relevant ways to determine such things as the relationships of the trainee scores to other scores and what improvements occur across training periods.

A comprehensive review of the literature bearing on approaches to the measurement of physician performance was authored by Arlene Barro (1973). Barro characterized the research which has been accomplished related to measurement of performance of physicians as falling into five major categories including: (1) the technical process approach, (2) the interpersonal process approach, (3) the simulated process approach, (4) patient outcome measures, and (5) the physician qualities approach. The last is that which has been primarily pursued in our Utah studies.

Our studies have been broad in coverage, not particularly differentiating between different types of physician specialties, but looking at the total problem of what is an excellent physician and how to select and train a person to be an excellent physician. The same broad approach has typically been taken by committees selecting applicants into Schools of Medicine.

Barro's general conclusions following her review were:

> The following general impressions of the state of the art of physician performance measurement emerged from this survey.
>
> First, at the present time, there exists no system for measuring the overall performance of individual physicians that has been validated in the sense that physicians who measure higher have been shown to produce better patient outcomes. In other words, no system of measurement now available allows us to determine objectively who are the high- and low-performing physicians. Lack of validation in terms of patient outcome is the primary shortcoming of many of the approaches currently being used or developed, and efforts to establish such validity should be the focus of future endeavors in this field.
>
> Regarding dimensions of performance, there has been a marked tendency to concentrate on the technical aspects of performance and to neglect the interpersonal . . . . There have been few attempts to define performance categories and to assign weights based on empirical analysis of the composition of actual practice. More effort in this direction would contribute to greater content validity of performance measurement instructions (p. 1085).

Intensive studies on outcomes typically and almost inevitably have to cover a smaller range of physician performances and characteristics than we have continually tried to deal with in our more comprehensive studies.

Thus, a great deal of research has been conducted here and elsewhere to define the "end product" of the medical profession, namely the practicing physician, and to use this as a clear-cut target against which to select and develop the forthcoming members of the medical profession. It was intended that the criterion measures developed for practicing physicians would have sufficient generality to permit suitable modification for the earlier-stage assessment of the performance characteristics of residents, interns, medical students, and pre-medical students.

IV. RESULTS

A. Physicians in Practice. The items across the five forms were classified according to their nature into conceptually and empirically homogeneous clusters. This yielded the 11 following types of conceptual scores: (1) Diagnostic-Therapeutic Knowledge, Skills, and Performance; (2) Keeping Abreast; (3) Effectiveness in Communicating with and Instructing Others; (4) Quality of Business Practices; (5) Human Relations Qualities in Working with Other Health Professionals (Physicians, Nurses, Technicians); (6) General Human Relations Qualities in Dealing with Patients; (7) General Intellectual Characteristics; (8) Physical Endurance; (9) Ethics and Character; (10) Altruistic Dedication to Medicine; and, (11) Characteristics that are Negative for Physicians.

These different types of scores were analyzed per source according to good psychometric attributes. Thus we computed each appropriate type of reliability and validity coefficient that was feasible. For example, where the data on scores per source permitted, internal consistency types of reliability coefficients were determined. Nearly half of these coefficients were about .90 or above, even though the number of items per score was often quite small.

Where a given type of score was obtainable from two or more of the five sources, a total score combined across sources becomes possible. If each type of conceptual score per source were combined across sources, the estimated reliabilities for these composite scores (lengthened by adding the item scores across the two to five sources) would be higher, with at least eight out of the 11 reliability coefficients becoming about .90, or higher.

With few exceptions, the scores were found to meet more than adequately the technical measurement characteristics desired, such as item reliability, item validity, composite score reliability, and composite score validity. These generally positive results on psychometric findings were the main outcome in this study.

B. Interns and Residents in Training. Early in the development of the trainee evaluation instruments, 3 subscores were identified through factor analytic techniques. These subscores were (1) Surgical Knowledge and Skill, (2) Ability to Work with Others, and (3) Desirable Physician Characteristics. Each subscore consisted of 10, 8, and 5 items, respectively. An overall rating and a commendation-retention-dismissal item were included and averaged across raters, with all written comments being added to the feedback.

The ratings for the knowledge-skill items increase significantly as the trainees progress through the training program. However, the other two subscores show little or no significant gain, suggesting that little real training is occurring in those physician-like attributes during residency training. Nonetheless, observer-raters are able to measure these other two scores even though the scores do not particularly improve during the training received from these observer-raters.

Correlations were obtained between the three subscores and the in-service tests (The American College of Surgeons In-Training Examination for General Surgery). The results indicate that only the knowledge - skill subscore was significantly related. This suggests that (a) there is significant overlap and (b) the local evaluation forms are apparently getting at other, independent abilities missed by the national test.

V. DISCUSSION

This latest study of physician characteristics and performances was an expansion in coverage and an intensification in depth of our earlier study of 80 measures of the performance of a larger number of physicians. It demonstrated the feasibility and the positive psychometric qualities of the broader coverage of measures of physicians in practice.

The study of physicians-in-training illustrates the practicality of producing measures that are relevant, in this earlier stage of training, in the sense of being largely on the target of physicians-in-practice. This implementation effort also uncovered problems in a nation-wide specialty exam and in the need to learn better how to train interns and residents in the ability to work with others and in other physician-like characteristics.

Reflecting about these results brought to mind a preliminary study which compared the performances and characteristics displayed by intern-residents with those of practicing physicians. The results of the study indicated that 73 of the 85 positive characteristics (86%) of the intern-resident list were also present in the physician list and conversely, 73 of the 87 positive attributes (84%) present in the physicians list were also present in the intern-resident list. Residents still have not had a few physician-like opportunities such as running their own offices and having the full responsibility for patients and for their own practices (Price and Taylor, 1973).

The high percentage of about 85% overlapping attributes and performances between residents and the physicians contrasted sharply with this research team's earlier correlational findings of essentially zero relations between performances of medical students and performances of practicing physicians (Price et al, 1971).

VI. CONCLUSIONS AND IMPLICATIONS

The results of these studies are generally supportive of the criterion-focused approach and expectations that we have had of the feasibility of this criterion-measurement approach.

We have repeatedly pointed out the potential logical extensions of this large and continued effort to measure physician performance. These latest studies support and demonstrate these contentions. Using "the end product", the physician, as a reference target as students move through pre-medical, medical and specialty training, it would be expected that an increasing number of the component attributes of a practicing physician would be unearthed and developed. The earlier that these attributes are displayed (even in pre-medical years), the better the profession will be able to predict and select those who will be excellent physicians in the future rather than merely those who will be successful medical students. Furthermore with relevant, on-target measures throughout all these stages, it become possible for the lengthy educational programs to increase their power to produce graduates who will become excellent physicians.

The usefulness of the results of all these past studies can potentially be great. They can provide guidelines for pre-medical training and evluation committees, for selection committees, for curriculum development committees, and for career decisions of prospective students planning to enter medicine. A main practical outcome reported herein is that all these previous studies provide the basis for moving toward the construction of a set of instruments which provide on-target measures at all stages in the development of professional physicians. Intensive efforts are needed to construct and revise and revise again measures of performance of students in pre-medical and medical education programs and performance of physicians in-training in various specialties.

VII. REFERENCES

Albo, D. Jr., Taylor, C. W., Page, B., Chang, F. C., Moody, F. G. Multifactor evaluations of surgical trainees and teaching services. *Surgery*, July, 1976, *80*: 115-121.

Barro, A. R. Survey and evaluation of approaches of physician measurement. *Journal of Medical Education*, 1973, *48*: 1048-1093.

Gough, H. G. How to select medical students. *Medical Teacher*, 1979, *1*: 17-20.

Price, P. B., Taylor, C. W., Nelson, D. E., Lewis, E. G., Loughmiller, G. C., Mathiesen, R., Murray, S L., & Maxwell, J. G. *Measurement and prediction of physician performance*. Salt Lake City: Aaron Press, 1971. Now available from Calvin W, Taylor, Building 404, University of Utah, Salt Lake City, Utah 84112.

Price, P. B., Lewis, E. G., Loughmiller, G. E., Nelson, D. E., Murray, S. L., and Taylor, C. W. Attributes of a Good Practicing Physician. *Journal of Medical Education*, *46*: 229-237, 1971.

Taylor, C. W., Nelson, D. E., and Price, P. B. *Comprehensive Analysis of Physician & Physician In-Training Performance* (Vols. I and II). Report to Public Health Service, Contract #HSM 110 71 171, University of Utah, Salt Lake City, Utah, 1974.

FAILURE OF EXAMINATION ANSWERS TO EVALUATE ACTUAL PRACTICE PATTERNS BY MEDICAL HOUSESTAFF IN AN OUTPATIENT CLINIC

Steven D. Rose, M.D., Lourdes C. Corman, M.D., John Robbins, M.D.
Section of General Medicine, University of California Davis

Introduction

Much of the evaluation of medical education and training is done by formal testing and by discussion among attendings, housestaff and students during hospital rounds. It is not known if these means of evaluation reflect actual practice. This study is designed to find any discrepancies there may be between answers given by housestaff to questions concerning patient care and the housestaff's actual practice in the outpatient medicine clinic. The use of screening tests in new patients in a general internal medicine clinic was analyzed because of the straightforward manner one could ask examination questions, identify appropriate patients and obtain data from the charts to compare with the examination answers. The hypothesis was that a physician who, on examination identified a screening test as appropriate for all new medicine clinic patients would order the test in actual practice. The results of the test would be recorded in the chart by either the physician, the nurse or with a laboratory slip.

Methods

Ten second year residents, in their final six months of training in internal medicine at the University of California Davis Medical Center, were given an oral exam by either direct interview or over the telephone. Each resident was told to answer the questions as if they were taking their Boards in Internal Medicine. The questions were:

1. What part of the physical exam and what laboratory tests should be done for screening in all new adult patients in a primary care clinic? Assume the patients are asymptomatic in the area that the test evaluates.

2. In addition to your answer to No. 1, which of the following tests should also be done?

   a. Blood pressure
   b. Breast exam
   c. Pap smear
   d. Stool guaiac
   e. Cervical culture of gonococcus
   f. Urinalysis
   g. CBC

Requests for reprints should be addressed to Section of General Medicine, Room 205, Professional Building, 4301 X Street, University of California Davis Medical Center, Sacramento, California 95817

h. SMAC 20
i. VDRL
j. Chest x-ray
k. EKG
l. PPD

3. Do you actually order the tests to which you gave a positive answer in the above question in all of your medicine clinic patients?

The first question was meant to be open-ended for the resident to volunteer answers. The second question was to get responses to 12 screening tests that we wanted to analyze. The third question was to see if the residents themselves were aware of any discrepancy between their stated answers and their actual practice.

After the examination, charts of new patients, seen 1-5 months previously by each resident, were reviewed to find if the housestaff had done the tests they claimed they would do. Charts were excluded from the study if the patient had had a recent workup by another physician or if the patient was sent to the clinic for a consultation. All charts included were of patients not under the care of another primary care physician and who had little or no previous evaluation. Charts were reviewed to determine which screening test from the above list were ordered, had been done earlier or a note had been made. This data was then matched with the resident's responses. Only positive answers were analyzed. Tests which residents would not obtain were not compared to actual practice. Such negative answers could not be easily analyzed for there may have been any number of reasons not obvious from the chart as to why the test was ordered. The number of tests answered positively that were actually done were then divided by the total number of new patients seen by the residents.

Results

The results of the exam by the 10 residents are shown in Table I below. The first column represents the number of positive answers given by residents to the open-ended first question and thus are answers recalled by the physician without the benefit of a list to choose from. The second column represents the number of positive answers given to the second question and thus are positive multiple choice answers. Any increase in positive answers from the first to the second column represents tests that the residents did not recall in question number one as correct but did so when presented with the multiple choice list in question number two. The third column represents answers to the third question and is the physician's estimate of what he/she actually orders in his or her own patients. The fourth column represents the percentage of tests answered positively by question number two that were identified by chart review as being obtained. Blood pressure was not analyzed in that this test is done on every patient by the nursing staff.

Using the answers to question number two, the residents overall did 58% of all tests that they had indicated should be done on all patients. There was a large variation among individual residents ranging from 24%-88%. There was a particularly low percentage for the VDRL and the PPD. None of the three residents who identified a cervical culture for gonococcus as an appropriate screening test, actually obtained this test on any of their appropriate patients. Although other tests fared better, most did not get above 75%. A breast exam was noted in the chart in just over half of the appropriate patients.

TABLE I. STATED PRACTICES VS ACTUAL PERFORMANCE

(N = 10 residents)

	Test volunteered to question #1	Test chosen in question #2	Resident's estimate of actual performance	% Tests done confirmed by chart review
Blood Pressure	8	10	10	-
Breast Exam	6	9	9	54%
Pap Smear	7	10	10	68%
Stool Guaiac	8	10	10	64%
Cervical G C Culture*	0	3	1	0%
Urinalysis	5	5	5	71%
CBC	6	6	6	78%
Chem Panel (SMAC 20)	5	5	6	83%
VDRL	4	10	4	19%
Chest X-ray**	4	4	4	100%
EKG***	6	6	6	71%
PPD	3	8	3	18%

* 2 qualified 40 yrs.
** 1 qualified for smokers
*** 2 qualified 40 yrs.

Discussion

The large discrepancy between the positive answers identified on the list of screening tests in question number two and the number of tests actually recorded in the clinic rejects the hypothesis that testing reflects performance. The highest score for any test was only 83% and the average was 54% with some below 20%.

There is reason to believe the discrepancy is even larger. We could not determine easily which tests were ordered for the purpose of screening and which were ordered for the evaluation of symptoms. None of the patients reviewed had presentations such as genital lesions, rashes, dementia and findings of neuro syphillis or any organic aneurysms that would require any VDRL. Thus the VDRLs ordered were likely to be real screens. In contrast, tests such as the chest X-ray, EKG, CBC and SMAC 20 are often obtained because of symptoms or abnormal physical findings. Many of the patients in the clinic presented with fatigue and thus a CBC may be ordered to evaluate this. Indeed, 65% of the patients examined by residents who did not consider a CBC a screening test obtained a CBC. Residents who also claimed the use of CBC as a screening test obtained the test in 78% of their patients. It is possible that if patients did not present with what the physician thought were indications for a CBC the percentage of CBCs done may be lower than 78%. Similar arguments can be made for the chemistry panel, the chest X-ray and the EKG.

In a previous study, Goran, et al[1] examined a similar hypothesis using a patient management problem (PMP). They found that the performance of clinic physicians with patients simulated on a PMP did not reflect performance with similar real patients in clinic. Examining patients with abdominal and/or flank pain, clinic physicians obtained more history appropriate for evaluating a urinary tract infection on the PMPs than in the clinic. Likewise, appropriate cultures were ordered in 39% of real patients versus 85% of simulated patients. Furthermore, a study team examining the same real patients before the exam by the clinic physicians evaluated with each patient the history and physical items of the PMP. They found that 82% of these real patients had indications for urine culture. This resulted in nine positive cultures. In contrast, the clinic physicians had ordered half as many cultures as the study team resulting in only three positive cultures. Their conclusions were that performance on a PMP did not correlate with performance in real life.

The results of our study shares similarities and differences with Goran's findings. We also found that performance on examination did not correlate with performance in real life. Likewise, a large variation was found among the different items. Many patients were asked for a history of dysuria but none were asked for a history of catheterization. Similarly, we found that 18% of appropriate PPDs were done in contrast to all chest X-rays being ordered by residents after identifying this as a screening test. In contrast to Goran's study which examined the validity of a PMP, the present study tested knowledge by asking the same question in two different manners. The first question required recall of the screening tests. The second question provided a list from which one could check the appropriate answers. A PMP is much like our question number two in that it provides only lists and examinees can react to specific items without having to recall themselves these items.

This last difference may be important. PPDs and VDRLs which were both ordered in less than 20% of appropriate patients also were recalled by less than half of the residents that identified positively the screening tests when presented on a list of possible answers by the majority of residents. Tests recalled more often in question number one such as the EKG and chest X-ray fared somewhat better. A possible explanation is that in real life physicians act from recalling to mind the appropriate line of action and are not in the position to pick various lines of action from a list. In a PMP or multiple choice exam, physicians can correctly react to items they may not have thought of on their own to either act or not act upon. Indeed, if one analyzes only the patients of the residents who answered positively to the VDRL and the PPD in question number one, a higher percentage for each test was found. This however is not the total answer as the percentage rises only to approximately 40%.

The reasons for the discrepancy are not obvious. Possible explanations include the artificial situation of multiple item lists as discussed above.

Additionally, a resident may "know" a correct answer but not believe in it. The VDRL has been often promoted as an ideal screening test in the literature[3] and thus physicians may think the examiners are testing for knowledge of this literature even though the examinees don't believe themselves the test is worthwhile. Another possible cause is inconvenience. Busy medicine clinic schedules may preclude difficult screening tests for more important symptom oriented evaluation. However, tests like Pap smears, which require time, equipment and an attendant to be obtained fared better than much simpler tests that residents could simply order. Less likely explanations include the inability to obtain a valid evaluation of performance by chart review. This study was designed to make this a small factor. Credit was given either in ordering or if the result was found. Credit might not be given in a rare case if a test was ordered and not noted on the chart and further, the result was lost. However, a lost result with no follow-up will have the same effect as no test being ordered at all. Some patients may have been deemed appropriate for screening by chart review but not by the resident. We think this problem is minimal in that patients studied had no primary care doctor before presentation and came to clinic for a general evaluation. All patients had histories that addressed the major organ systems and a past medical history. Physicals were nearly complete except for breast exams which were left out of otherwise complete physicals. Evaluations were general and none were limited to specific problems.

There have been many debates on the different ways to test physicians and medical students.[4] Testing may look for knowledge or various abilities, such as the ability to solve problems.[2] Although it is recognized that there may be no correlation between testing and performance the degree of discrepancy has not been well studied. This study has disproved the hypothesis that a good correlation exists between knowledge and performance. There is a wide variation in performance among the residents tested as well as a wide variation among the specific screening tests. This indicates that further work must be done to evaluate performance in medical education. This study used chart review for evaluation. This method can only be used when very simple questions are asked that are well-defined and can be evaluated by easily retrievable data from the chart. More complex problems would not lend themselves to chart review.

Evaluation of performance with such problems may require direct observation. Formal testing by oral, written or computer examination should concentrate on questions that require recall of answers rather than giving multiple choice answers or choosing stated lines of action.

REFERENCES

1. Gorman, M.J., Williamson, J.W., Gonnella, J.S. The Validity of Patient Management Problems J. Med Ed. 48: 171 1973
2. Ebel, R.L. Ability Versus Knowledge in Testing Educational Achievement National Board Examiner 16: No. 8 1969
3. Frame, P.S., Carlson, S.J. A Critical Review of Periodic Health Screening Using Specific Screening Criteria Part 3: Selected Diseases of the Genitourinary System J. Family Practice 2: 189 1975
4. Barro, A.R. Survey and Evaluation of Approaches to Physician Performance Measurement J. Med. Ed. 48: 1048 1973

RELIABILITY IN FOLLOW-UP EVALUATIONS OF GRADUATES AS RESIDENTS*

Rosalia E. A. Paiva, Ph.D.
Southern Illinois University School of Medicine

Assessment of performance of medical school graduates during residency training provides a direct indicator of quality of the undergraduate medical curriculum. Also, clinical performance at the graduate level provides an intermediate set of criteria--in lieu of ultimate physician performance--for the validation of measures utilized by clinical faculty. Medical schools have traditionally subscribed to these views, but often have focused only on the assessment of performance as measured by Part III of the NBME examination. Not until recently have efforts been undertaken to systematically follow graduates into residencies to gather information on their performance "on the job." The importance of this information cannot be overstressed, particularly when considering: a) repeated findings of poor to almost non-existent relationship between written examinations and performance on the wards with real patients; and b) potential uses of these data for evaluation of the undergraduate clinical curriculum.

For residency follow-up assessment to serve this end, data collected must meet certain standards of quality so that valid conclusions can be drawn. These considerations have been a primary concern in planning and implementing the Southern Illinois University School of Medicine (SIU-SM) Residency Follow-up Study. The aim of the study is to determine SIU-SM's achievement of one major institutional goal, namely, to provide solid academic and clinical preparation for residency training. The purpose of this paper is twofold: a) to describe the project as a useful model for medical school curriculum evaluation and b) to report on a reliability study of residency supervisor ratings.

THE SIU RESIDENCY FOLLOW-UP PROJECT

Objectives

The specific objectives of this ongoing project are: a) to assess the competence of SIU-SM graduates as determined by performance during residency; and b) to obtain reflective feedback from graduates regarding the adequacy of the SIU-SM curriculum in preparing them for residency training.

Method

Instrument - The project was designed by clinical and educational faculty. Each step in the design was guided by recommended psychometric and test development procedures. The major effort centered on development of the instrument, a rating scale, to measure the clinical performance of the graduates. To assure validity of the data, a task analysis was performed and multiple sources consulted resulting in identification of 13 traits in a construct dealing with clinical skills, cognitive variables and non-cognitive attributes (See Table 1). To assure both validity and reliability, each trait was defined operationally and from two to four descriptors or components of behaviors were added to each. Four other items representing broader, more comprehensive, ratings were included. They are competence in medicine in general, competence in the chosen specialty, comparison with other house staff, and desirability as personal physician. A ten-point scale ranging from "outstanding" to "unsatisfactory" was used.

---

*Requests for reprints should be addressed to Rosalia E. A. Paiva, Ph.D., Office of Curriculum Affairs and Educational Resources, Southern Illinois University School of Medicine, P.O. Box 3926, Springfield, Illinois 62708

Two forms were prepared, one for self-evaluation, the other for evaluation of graduates by supervisors. The clinical performance items are identical; there is one additional item in the supervisor's instrument: "desirability of the graduate as personal physician." The forms also ask for comments regarding major strengths and weaknesses of the resident taking into account the operational definitions and descriptors for each variable.

Residency directors are requested to have the two supervisors most familiar with the resident's work complete one evaluation form each. However, because the data collection starts approximately midway through the first year of residency when some supervisors may have had insufficient opportunity for meaningful interaction with the resident, two additional categories are included in the scale on their form. These categories allow the rater to indicate whether the rating given is "based on uncertain information" or no rating is given because the evaluator has "no information available."

Supervisors are asked to rate residents on how they compare to the "ideal" first-year resident on all the descriptors for each trait. Graduates are asked to rate themselves based on how well they were prepared for residency training.

Direct curriculum evaluation data also are collected from graduates. They are asked to rate and comment retrospectively on all major segments of the SIU-SM curriculum and to base their ratings on the adequacy of their undergraduate training in preparing them for residency.

Data Collection - In preparation for the data collection and prior to graduation, students are asked to sign a release form authorizing SIU-SM to obtain information from their future supervisors as well as their NBME III or FLEX scores. The first mailing is sent to residents and their supervisors six to nine months after graduation. It includes a letter explaining the project, the evaluation forms, and a copy of the signed release form. Reminder post cards and new packages of forms are then forwarded to non-respondents at two and five weeks after the first communication. Follow-up telephone calls are made when reminder post cards and letters prove unsuccessful.

Subjects - Follow-up of graduates began with graduation of the first class in 1975. The follow-up has been completed for the first three classes. Overall return rates for the 1975 (n=24), 1976 (n=43) and 1977 (n=47) graduates have been 83 percent for self-evaluations and 95 percent for supervisor evaluations.

Analysis - Some of the studies on performance ratings are: comparison between supervisor and self-ratings on all variables; relationship between ratings within and between groups; comparison of ratings across classes;

factor analysis of self and supervisor ratings; level of agreement between supervisors; agreement between supervisor and self-rating across variables for each graduate; and relationship between ratings and other data collected during medical school or after graduation.

Graduates' ratings of the curriculum are compiled per class, compared through classes and correlated with performance. Also, comments on the curriculum are studied in relation to curriculum evaluation data which graduates provided during medical school. Because graduates produce curriculum ratings and comments retrospectively, interpretation of the same includes consideration of changes and refinements that have occurred over time.

For each class surveyed to date, both graduates and their supervisors have provided extensive comments throughout the evaluation forms. A content analysis of comments is performed in order to identify categories of responses and trends related to specific strengths and weaknesses. Comments regarding both performance and curriculum are analyzed in this fashion. Data generated through this ongoing study are presented to faculty on a yearly basis in an annual report. The reports are distributed to all major curriculum committees.

Selected Results

1. Supervisor ratings are higher than self-ratings on most variables. The only expection are "history-taking ability" and "relationship with patients and their families." For these two variables graduates view themselves as significantly more competent than do their supervisors.

2. Supervisor ratings do not indicate differences among the classes.

3. There is no relationship between self and supervisor ratings on any variable except for "identification of problems/differential diagnosis" and "relationship with other medical personnel."

4. Total mean self-ratings show a strong relationship with graduates' ratings of major portions of the curriculum.

5. There seems to be a response bias on the part of graduates; non-respondent graduates obtained lower supervisor ratings on all variables but one than did graduates who completed their self-evaluations.

RELIABILITY STUDY

Because educational decisions and continuing program development should be at least partially based on the evaluation of outcomes of the instructional process, reliability studies constitute a necessary part of projects of this kind. Lack of generality of supervisor (or observer) ratings is a common finding in psychological and educational research. Multiple sources of error and bias can affect the reliability of observational measures, which in turn will restrict the validity of the data. Poor interrater agreement is often a major source of error. Careful methodological design and instrument development will reduce the contribution of error to the observed measures. However, because of limitations intrinsic to any measurement endeavor, and "because the datum recorded on a rating scale is the subjective judgment of the rater, the generality of a set of ratings is always of concern (Tinsley and Weiss, 1975, p. 358). Accordingly, the purpose of this study is to assess the quality of the supervisor ratings obtained in the SIU-SM Residency Follow-up Project and thus determine the extent to which the results are generalizable.

Method

Subjects - Subjects were all graduates who were rated by two residency supervisors. They include 13, 32 and 34 subjects in the classes of 1975 through 1977 respectively.

Analysis - Studies of the reliability of observational measures often report either interrater agreement or interrater reliability. Interrater agreement indicates the degree to which raters tend to assign the same rating to the subject they are judging. Interrater reliability is a measure of the degree to which two or more judges rank subjects in the same relative order. These two types of coefficients provide different kinds of information. It is possible for raters to show perfect agreement and yet produce very unreliable data. This occurs when there is little variability in the ratings or the trait judged is not stable. The opposite case is also true, that is, very high reliability but no agreement. To present a more complete and accurate picture of the relationship between ratings, it is suggested that both types of coefficients be reported (Tinsley and Weiss, 1975; Mitchell, 1979).

Interrater agreement was measured on each variable for each class by Lawlis and Lu's chi-square index (1972) followed by Tinsley and Weiss' T index (1975). Lawlis and Lu's index is appropriate for interval scales, permits the definition of agreement to include small discrepancies, and indicates whether the observed agreement is significantly larger than chance agreement. For this study, agreement between raters was defined as a discrepancy of one scale point or less on a ten-point scale. Whenever the Lawlis and Lu's chi-square index was significant, the Tinsley and Weiss' T index was calculated to determine the strength (high, medium or low) of the agreement.

Interrater reliability was determined for each variable by intraclass correlation indices using an analysis of variance approach (Shrout and Fleiss, 1979). Because each graduate was rated by a different pair of supervisors, a one-way analysis of variance with supervisors considered as a random sample from the population of supervisors was used. Two intraclass correlations indices were obtained; ICC (1,1) which gives the expected reliability of a single supervisor's ratings, and ICC (1,2) which gives the reliability index for mean ratings over two supervisors. Provided ICC (1,1) is positive, the ICC (1,2) index is always larger than the single-rater reliability index.

All analyses were performed separately for each class.

Results

Table 1 presents the two indices of interrater agreement on each variable for the three classes. For clarity, the chi-square values have been replaced with their levels of significance. Table 2 presents the single rating and mean rating reliability indices on each variable for each class.

Overall, the results show that supervisors a) tend to assign the same rating to the same graduate, b) differentiate among graduates, and c) rate them in a similar relative order for most variables. In regard to interrater agreement (Table 1), supervisors of the class of 1975 had the highest level of interrater agreement and those for the class of 1977 the lowest.

Concerning interrater reliability (Table 2), the results for the class of 1975 show lack of reliability on four variables: "physical examination skills." The variability of supervisor ratings was smallest for these variables. However, because interrater agreement was moderate (Tinsley and Weiss' T ranges from .47 to .65), it is possible that the rated graduates were homogeneous on these variables. A similar observation applies to "independent learning habits" for the class of 1977. This is the only variable that showed non-significant interrater reliability for the classes of 1976 and 1977. Inspecting intraclass correlation indices across classes shows that ratings on the narrowly defined variables (#1 through 13) were as reliable as those on the broad, multidimensional variables (#14 through 18).

DISCUSSION AND CONCLUSIONS

The results of the reliability study strongly indicate that the instrument developed for the Residency Follow-up Project is reliable, and reinforce the need for careful construction, operational definitions, and behavioral anchors.

The results of the reliability study should be interpreted in relation to the number of graduates in the study; they represent 54%, 74%, and 72% of the three classes. The Class of 1975 contributed the lowest percentage and the lowest absolute number (12) of subjects. Out of the 15 variables with available data for this class, four failed to show interrater reliability. The same variables had significant interrater reliability in the other two classes. The effect of number of subjects on variability of scores and therefore on reliability estimates is well known.

Establishing the reliability of residency supervisor ratings has been an important step in the continuing development of the Residency Follow-up Project. Indeed, the results of the reliability study have provided the basis for justifying the use of the data for making generalizations, drawing conclusions and decision-making. The next main step in the project deals with validation, especially in the area of construct validity.

REFERENCES

Lawlis, G.F., & Lu, E. Judgment of Counseling Process: Reliability, Agreement, and Error. Psychol. Bull., 78: 17-20, 1972.

Mitchell, S.K. Interobserver Agreement, Reliability, and Generalizability of Data Collected in Observational Studies. Psychol. Bull., 86: 376-390, 1979.

Shrout, P.E., & Fleiss, J.L. Intraclass Correlations: Uses in Assessing Rater Reliability. Psychol. Bull., 86: 420-428, 1979.

Tinsley, H.E.A., & Weiss, D.J. Interrater Reliability and Agreement of Subjective Judgments. J. Couns. Psychol., 22: 358-376, 1975.

TABLE 1

AGREEMENT BETWEEN SUPERVISORS ON EACH VARIABLE FOR THESE CLASSES†

		Lawlis & Lu's X^2			Tinsley & Weiss'T		
#	Variable	1975 (N=13)	1976 (N=32)	1977 (N=34)	1975 (N=13)	1976 (N=32)	1977 (N=34)
1	History	***	***	***	.89	.60	.41
2	Physical Examination	**	***	**	.54	.69	.33
3	Indentification of Problems	***	***	***	.68	.44	.45
4	Diagnostic Approach	***	***	***	.79	.82	.54
5	Management	***	***	***	1.00	.46	.49
6	Fund of Knowledge	***	**	NS	.79	.35	-
7	Independent Learning Habits	ND	***	***	ND	.60	.42
8	Relationship w/Patients & Family	NS	***	***	-	.68	.51
9	Relationahip w/Med. Personnel	**	***	***	.57	.52	.47
10	Reliability & Dependability	**	***	**	.47	.57	.35
11	Responsibility	***	NS	*	.68	-	.28
12	Emotional Stability	***	***	NS	.65	.48	-
13	Technical Skills	**	***	*	.57	.49	.31
14	Competence/Medicine	ND	***	***	ND	.64	.55
15	Competence/Specialty	ND	***	***	ND	.60	.66
16	Overall Competence	***	***	***	.79	.65	.51
17	Comparison With Other House Staff	***	***	***	.65	.40	.51
18	Desirability as personal physician	***	***	**	.83	.46	.36

*p<.05 **p<.01; ***p<.001

ND = No data; NS + Non-significant; "- = Not applicable

†Agreement defined as ratings differing by one scale point or less.

TABLE 2

INTERRATER RELIABILITY: SINGLE-RATER AND MEAN RATING INTRACLASS CORRELATION RELIABILITY INDICES ON EACH VARIABLE FOR THREE CLASSES

	Class of 1975 (N=13)		Class of 1976† (N=32)		Class of 1977 (N=34)	
# Variable	ICC(1,1)	ICC(1,2)	ICC(1,1)	ICC(1,2)	ICC(1,1)	ICC(1,2)
1	.80***	.89***	.44**	.61**	.53***	.69***
2	.38	.55	.61***	.76***	.44**	.61**
3	.69**	.82**	.54***	.70***	.41**	.58**
4	.69**	.82**	.59***	.74***	.54***	.70***
5	.94***	.97***	.45**	.62**	.61***	.76***
6	.83***	.91***	.42**	.59**	.35*	.52*
7	No Data	No Data	.74***	.85***	.27	.43
8	.48*	.65*	.49**	.66**	.60***	.75***
9	.50*	.67*	.57***	.73***	.64***	.78***
10	.41	.58	.53***	.69***	.64***	.78***
11	.59*	.74*	.38*	.55*	.56***	.72***
12	.32	.48	.63***	.77***	.45**	.62**
13	.38	.55	.46*	.63*	.40*	.57*
14	No Data	No Data	.63***	.77***	.63***	.77***
15	No Data	No Data	.47**	.64**	.69***	.82***
16	.75**	.86**	.57***	.73***	.67***	.80***
17	.71**	.83**	.43**	.60**	.68***	.81***
18	.85**	.92**	.59***	.74***	.63***	.77***
Total Mean Rating	.71**	.83**	.67***	.80***	.67***	.80***

*p<.05 **p<.01; ***p<.001

†For the Class of 1976, N was 21 for Technical Skills

THE EFFECTS OF CLINICALLY RELEVANT MULTIPLE-CHOICE ITEMS ON THE STATISTICAL DISCRIMINATION OF PHYSICIAN CLINICAL COMPETENCE

Steven M. Downing, Ph.D.
Jack L. Maatsch, Ph.D.
Office of Medical Education Research and Development
Michigan State University[1]

Purpose of This Study

The major purpose of this study was to investigate the effect of clinically relevant multiple-choice item content on the validity of discriminations of physician clinical competence. Research data were collected during a field test of the Emergency Medicine Examination (EME)--a multi-format examination ultimately intended for certification of specialists in Emergency Medicine.

The study compared the relative discriminating power of a multiple-choice subscale composed of items selected for high clinical-relevance to a subscale of items selected for middle difficulty. High clinical-relevance items were selected for their highest correlation with an independent rating of simulated clinical performance, while medium difficulty items were selected for their middle difficulty levels. Additionally, the internal-consistency reliabilities and the mean item difficulties of these two multiple-choice subscales were investigated.

Review of the Literature

During the past decade, increasing pressure has been brought on specialty certifying boards to demonstrate the validity of their measurement procedures for the prediction of the clinical competence of certified specialists. Several empirical studies, investigating the relationship between scores on certification examinations and measures of physician performance, have been reviewed by Williamson (1976). One of the earliest studies, the Teamster Study (Trussell, 1962), concluded that the certification status of physicians was, for the most part, unrelated to the quality of health care delivered. This study was replicated and the findings confirmed by Morehead and others (1964).

McGuire and Williamson (1968) studied the performance of general practitioners, non-certified, and certified physicians on patient management problems. No statistical difference in performance was observed among these three groups. Pawluk and others (1976) report low positive or negative correlations between various certification format scores and an independent measure of quality of care for Canadian Family Practitioners. Gonnella (1973) found low correlations between scores on multiple-choice examination scores and measures of physician diagnostic accuracy and proper patient management.

---

[1] This research was supported in part by a grant from the National Center for Health Services Research (HS02038). The data, test formats, scoring mechanisms and all related examination development and validation procedures described in this paper were developed for the American College of Emergency Physicians. The American Board of Emergency Medicine, which will subsequently administer the first certification examination, reserves the right to use all or part of the test library and methodologies developed by the American College of Emergency Physicians. Reprints: Steven M. Downing, Ph.D., Michigan State University, Office of Medical Education Research and Development, A-202 East Fee Hall, East Lansing, Michigan 48824.

Some of the best evidence for the lack of criterion-related validity for certification examinations was presented by Payne and Lyons (1972). From a large-scale study of the correlates to competent health-care delivery, these researchers concluded that board certification, type of specialization, years of practice, and hospital size did not correlate well with process audit ratings of physician performance.

On the other hand, Maatsch and others (1978) reported that both Multiple-Choice and Simulated Clinical Encounter scores are sensitive to years of training and experience in Emergency Medicine. Preliminary analyses of data from a field test of the EME showed that a total score on the objective format correlated .83 with independent ratings of clinical performance on realistic, patient-game simulations of typical emergency department clinical problems.

Methodology

This section presents a brief description of EME construction, the sample of subjects for this study, the experimental design, and the statistical methodology employed.

Instrument Development

The EME was developed by the American College of Emergency Physicians (ACEP) and the Office of Medical Education Research and Development (OMERAD), Michigan State University. This Examination consists of two objective-item formats (364 Pictorial Multiple-Choice and Multiple-Choice items), 12 Patient Management Problems (PMP), and 12 examiner-administered Simulated Clinical Encounters (SCE).

All Examination materials were field tested with ninety-four subjects who represented four distinct groups in training, experience, and ability to deliver health care in Emergency Medicine. These groups were: 1) Residency-Eligible Emergency Physicians: Subjects potentially eligible to take the EME by virtue of graduation from an approved residency in Emergency Medicine and at least one year of continuous practice in Emergency Medicine (n=22); 2) Practice-Eligible Emergency Physicians: Subjects potentially eligible for EME by virtue of five years of continuous practice in Emergency Medicine (n=14); 3) Emergency Medicine Residents: Residents beginning their second post-graduate year (n=36); and 4) Fourth-Year Medical Students: Volunteer subjects beginning the final year of pre-doctoral study (n=22).

Emergency physicians were selected by a national stratified random sampling procedure designed to select subjects judged by their peers as competent clinicians and who were, therefore, certifiable in Emergency Medicine. Residents in this study represented a national stratified random sample of all Emergency Medicine residency programs; subjects were selected for a wide-range of clinical competence. Medical students were paid volunteers.

All subjects completed all examination materials during two and one-half days of testing. Objective items were administered under standard testing conditions and the examiner-administered SCEs were presented and rated by well trained emergency physicians examiners. The interclass inter-rater reliability of the SCE grand mean rating was .79 (Maatsch et al., 1978). Group membership of individual subjects was unknown to SCE examiners.

Research Hypothesis

The major research hypothesis under investigation here is that an empirically defined multiple-choice subscale, consisting of high clinical-relevance items, will statistically separate residency-eligible physicians, residents, and students better than a multiple-choice subscale consisting of medium difficulty items.

The following subscales were selected for this investigation: 1) Medium Difficulty Subscale (MD): 91 items with ideal item p-values (proportion correct) for norm-referenced achievement tests (Ebel, 1972). P-values ranged from p=.50 to .69 with mean p=.603. All items had positive point-biserial item-total score discrimination indices; and 2) High Clinical-Relevance Subscale (HCR): 91 items with the highest item-mean SCE rating point-biserial correlations. Item-criterion correlations ranged from r=.33 to .68 with the median r=.38. The mean p-value for this scale was .721.

Results

TABLE I
RAW-SCORE GROUP DISCRIMINATION:
MEDIUM DIFFICULTY VS HIGH CLINICAL-RELEVANCE

	Residency-Eligible		Residents		Students		Wilks'	$F_{2,77}$
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Lambda	
High Clinical-Relevance	80.23	7.67	70.81	7.59	42.55	9.98	.2353	125.21*
Medium Difficulty	66.77	8.67	56.83	7.89	39.77	7.14	.3686	65.94*

*$p \leq .0001$

Table I presents the raw-score group discrimination of three criterion groups by each subscale. This analysis shows that each subscale statistically separates groups, since scale means are ranked in the predicted direction and each F-ratio is statistically significant. Wilks' Lambda, a multivariate index of the group discriminating power of a variable (Tatsuoka, 1971), indicates that the HCR scale accounts for more discriminating power than the MD scale. (A small Lambda indicates high discrimination power.)

TABLE 2
SUMMARY OF STEPWISE DISCRIMINANT ANALYSIS:
HIGH CLINICAL-RELEVANCE VS MEDIUM DIFFICULTY

Step Number	Scale Name	F to Enter	Wilks' Lambda
1	High Clinical Relevance	125.21	.235
2	Medium Difficulty	4.38	.211

Table 2 gives the results of a Stepwise Discriminant Analysis of these two scales for the three criterion groups in this study. This procedure selected the most discriminating variable first, entered this variable into the analysis, and then chose the second-best discriminating variable for entry into the analysis. From Table 2 it is apparent that the HCR scale is considerably more discriminating of known groups than the MD scale, since there is little change in Lambda from step one to step two in this analysis.

Two discriminant functions were derived by the Stepwise Discriminant Analysis. In the first function--the HCR function--the standardized discriminant function coefficient (2.374) was about 6.7 times more important than MD (-.357) to the separation of known groups.

TABLE 3
CLASSIFICATION ANALYSIS USING HIGH
CLINICAL-RELEVANCE AND MEDIUM DIFFICULTY DISCRIMINANT FUNCTIONS

			Predicted[a]	
	N	Residency-Eligible	Residents	Students
Residency-Eligible	22	15(68.2)	7(31.8)	0(0)
Residents	36	6(16.7)	29(80.6)	1(2.8)
Students	22	0(0)	1(4.5)	21(95.5)

---

[a] Number in parentheses indicates percentage of classification for that group.

Table 3 presents a classification analysis of the predicted group membership versus the actual group membership, using the two discriminant functions extracted by this analysis. A total of 81.3 percent of subjects were correctly classified by both discriminant functions--HCR and MD. A chi-square test statistic ($\chi^2=82.66$, df=4, $p \leq .0001$) was computed to test an hypothesis of chance classification accuracy. The high clinical-relevance scale, taken separately, classified 76.3 percent of subjects in known groups correctly, while the medium difficulty scale classified only 71.2 percent of cases correctly.

Subscale Reliability

Internal-consistency reliability was calculated for each 91-item subscale for the N=80 subjects in this study. The KR20 r for the HCR scale was .954, while r=.878 for the MD scale. An F-test of this observed difference in scale reliability is given, following Hoyt (1941) and shown by Downing (1979), by a ratio of the Mean Squares associated with the reliability calculation for each subscale. For this null hypothesis, the F-test (F=2.06, df=79, 79, $p \leq .05$) showed that the HCR subscale was statistically higher in internal-consistency reliability than the MD subscale.

Item Difficulty

TABLE 4
SUBSCALE MEAN ITEM DIFFICULTY
N=80

Subscale	Mean	Standard Deviation	Mean P-Value	Z-Value
High Clinical-Relevance	65.63	16.96	.721	13.48*
Medium Difficulty	54.86	12.88	.603	

$p \leq .0001$

Table 4 gives the mean item difficulty for each scale. A correlated Z-test of the difference in scale means shows that the HCR scale is significantly less difficult than the MD scale.

Conclusions and Discussion

The following conclusions follow from the results of this study:

1. Multiple-choice items can be written and calibrated to be powerfully valid discriminators of varying levels of clinical competence.
2. Selection of multiple-choice items for their high correlation with a valid and reliable independent measure of clinical performance is more efficient for the valid discrimination of physician clinical competence than selection of items for a middle difficulty criterion.
3. Selection of multiple-choice items for their high correlation to a simulated clinical performance measure yields a more reliable measurement than selection of items for their middle difficulty.
4. Clinically relevant items are less difficult for clinicians than items typically used for certification or licensure.

These findings must be viewed as tentative and generalizations to other settings must be cautious, since subjects for this study may be atypical of subjects in other certification settings. However, the results of this study do suggest that the relevance of multiple-choice item content to the typical practice of clinical medicine is important to the valid statistical discrimination of levels of physician competence.

It is also interesting to note that high clinical-relevance items tend to be less difficult than typical, norm-referenced achievement items. Yet, the less difficult clinically relevant items yield highly reliable measurements of physician competence.

If these results generalize to other specialty certification settings, certifying boards would do well to select objective-item content that is maximally related to the typical and competent practice of clinical medicine.

Further research is indicated to replicate these findings, using larger samples of subjects. Validation and cross-validation studies are also essential. Another important area of research suggested by this study is the determination of the content selection rules for HCR scales.

References

Downing, S.M. "An Analysis of the Effects of Different Multiple-Choice Item Selection Strategies on the Reliability and Validity of Measures of Physician Competence in Specialty Certification," 1979, A Dissertation at Michigan State University.

Ebel, R.L. Essentials of Educational Measurement, Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1972.

"Emergency Medicine Condition/Skills List," Journal of the American College of Emergency Physicians, August, 1976.

Gonnella, J.S. "Evaluation of Patient Care: An Approach," Journal of the American Medical Association, Vol. 214, No. 11, 1973, pp. 2040-3043.

Hoyt, C.J. "Test Reliability Estimated by Analysis of Variance," in Principles of Educational and Psychological Measurement, eds. W.A. Mehrens and R.L. Ebel, Chicago, Illinois: Rand McNally and Company, 1967.

McGuire, C.; Williamson, J.W. "Consecutive Case Conference: An Educational Evaluation," Journal of the American Medical Association, Vol. 43, No. 10, 1968, pp. 1068-1074.

Maatsch, J.L.; et al. "Toward a Testable Theory of Physician Comeptence: An Experimental Analysis of a Criterion-Referenced Specialty Certification Test Library," in Proceedings of the Seventeenth Annual Conference on Research in Medical Education, Washington, D.C.: Association of American Medical Colleges, 1978.

Morehead, M.A.; Trussell, R.E.; Ehrlich, J. A Study of the Quality of Hospital Care Secured by a Sample of Teamster Families in New York City. New York: Columbia University Press, 1964.

Pawluk, W.; et al. "Concurrent Validity of the Canadian Certification Examination in Family Medicine," in Proceedings of the Fifteenth Annual Conference on Research in Medical Education, Washington, D.C.: Association of American Medical Colleges, 1976.

Payne, B.C.; Lyons, T.F. Method of Evaluation and Improving Personal Medical Care Quality, Office Care Study for Hawaii Medical Association, Ann Arbor, Michigan: University of Michigan School of Medicine, 1972.

Tatsuoka, M.M. Multivariate Analysis: Techniques for Educational and Psychological Research, New York: John Wiley and Sons, Inc., 1971.

Trussell, R.E. The Quantity, Quality and Costs of Medical and Hospital Care Secured by a Sample of Teamster Families in the New York Area, New York: Columbia University Press, 1962.

Williamson, J.W. "Validation by Performance Measures," in Conference on Extending the Validity of Certification, Chicago, Illinois: The American Board of Medical Specialties, 1976.

PRECIS

PREDICTION AND SELECTION

MODERATOR: Paul R. Elliott, Ph.D.
Florida State University

THE RELATIONSHIP BETWEEN SCORES ON THE NEW MEDICAL COLLEGE ADMISSION TEST AND PERFORMANCE IN THE FIRST YEAR OF MEDICAL SCHOOL: A PRELIMINARY STUDY

This paper examines the relationship between performance on the subparts of the New MCAT and performance on examinations administered to first year medical students. The results indicate at least a moderate degere of correlation between the New MCAT and first year examinations.

A FACTOR ANALYTIC STUDY OF THE OLD AND NEW MCAT EXAMINATIONS

The old and new MCAT examinations were factor analyzed independently, each yielding a two-factor solution. The structure of the new MCAT was compared with the old test, and recommendations for using the new test are presented.

A NON-LINEAR DISCRIMINANT ANALYTIC APPROACH TO PREDICTING MEDICAL STUDENT PERFORMANCE

The authors present a method for predicting medical student performance considered potentially useful to the admissions process. The method, which employs discriminant function analysis augmented by explicit consideration of non-linear relationships, is validated through retrospective study of three classes of medical students.

A STRUCTURED INTERVIEW FOR THE SELECTION OF PHYSICIAN'S ASSISTANT STUDENTS

The structure of a personal interview for the selection of physician's assistant students is described. Results supported the inter-rater reliability and concurrent validity of the interview.

The Relationship Between Scores on the New Medical College Admission Test and Performance in the First Year of Medical School: A Preliminary Study

Bruce C. Deighton, Douglas U. Smith, Ph.D., Richard E. Gallagher, Ph.D.[1]

The New Medical College Admission Test (New MCAT) was designed to upgrade the quality of information concerning a candidate's potential for success in medical school and medical practice. The New MCAT replaces the Medical College Admission Test which has been in use since the late 1940's. Whether the New MCAT represents a significant improvement in the medical school admissions process is a question of current interest to admissions committees and others concerned with medical education.

BACKGROUND

Historically, the Medical College Admission Test has been used as a predictor of academic performance in the early, or basic science, years of medical training. Since most dropouts occurred during the first two years of medical school, MCAT scores were an important factor in the selection of candidates who would successfully complete the course of training and go into professional practice.

Several studies, however, questioned the effectiveness of the MCAT in performing this task (Watson, 1955; Garfield and Wolpin, 1961: Fredericks and Mundy, 1967; Bartlett, 1967; Howell and Vincent, 1967). Richards (1972) in a review of the MCAT published in Buros' Seventh Mental Measurement Yearbook reported correlations between MCAT subtests and rank in class to range from .06 to .22 for the pre-clinical years, and from .09 to .14 for the clinical years. Gough, Hall, and Harris (1963) in reviewing research concerning medical school admissions, noted that the validity of the MCAT in predicting academic performance and completion of training was low and when used to predict performance in postgraduate training and professional practice was "essentially zero."

Admissions committees were faced with the problem of evaluating candidates in terms of MCAT scores which apparently did not differentiate well between success and failure in medical school, nor contribute to the prediction of performance following graduation. However, as Gough et al. (1963) pointed out, many medical educators viewed the MCAT as having "intrinsic validity" in selecting candidates who had qualities desired in a professional practitioner. At the same time, the notions as to the qualities desired in practicing physicians came under increasing examination, (e.g., Haley, Juan and Paiva, 1971; Erdmann, 1972). Medical educators were concerned that the admissions process tended to select individuals who would become physicians who were technically competent, but perhaps lacking in humanistic qualities desired in physicians, (Ceithaml, 1976, p. 34).

---

[1] Requests for reprints should be addressed to Richard E. Gallagher, Ph.D. Associate Professor and Director, Division of Educational Services and Research, Wayne State University School of Medicine, Detroit, Michigan 48201

In 1972, a national study group of the Association of American Medical Colleges undertook the task of evaluating admissions procedures in use at that time and recommending necessary changes. The need for a major revision of the Medical College Admission Test was identified, resulting in the construction of the New MCAT. The New MCAT ostensibly provides a broader basis of information, more relevant to medical school curricula and medical practice. Based on this consideration, as well as the use of updated testing methodologies, the predictive validity of the admission test should improve.

Preliminary testing conducted by the AAMC concurrent with the development of the New MCAT showed moderate to high correlations between New MCAT scores and academic performance in medical school (New Medical College Admission Test Interpretive Manual, 1977). These studies were subject to a number of concerns, such as restricted variability on both criterion and predictor variables, unknown reliabilities of criterion measures, and uncertainty regarding the motivation and preparation of the sample populations, which tended to limit the generality and utility of the results. As far as can be determined, the AAMC studies are the only reports on the validity of the New MCAT to date.

The present study is an attempt to examine the relationship between scores on the New MCAT subtests and performance on first year unit examinations for the 1978 entering class at a large, urban, midwestern medical school.

METHOD

A total of 252 first year students were considered in the present analysis. Four students repeating the first year were excluded, since they did not have New MCAT scores. The composition of the class included 191 (75.8%) males, and 61 (24.2%) females. The average undergraduate grade point average was 3.58/4.00 -- 3.56 in the sciences, and 3.60 in all other areas.

A multidisciplinary teaching approach is employed in Year I with content organized into seven instructional units using an organ system approach. In Year I these units are entitled: Introduction (INTRO), Cardiovascular (CV), Gastrointestinal (GI), Respiratory/Renal (RESP/RENAL), Endocrine (ENDO), Neuroscience (NEURO), and Family Medicine (FMED). Year I curriculum deals with the normal structure and function of man. The second year program seeks to develop an understanding of the basic pathophysiology of the common diseases of man.

Evaluation of student performance was based on multiple-choice type examinations administered at the end of each unit of instruction. Items comprising the examinations reflected the amount of instructional time allocated to each topic. A comprehensive examination covering material from each of the seven units is administered at the end of each academic year.

Simple correlations between each of the New MCAT subtests: Biology, (BIOL), Chemistry (CHEM), and Physics (PHYS) knowledge; Science Problems (PROB); Reading (SA/R) and Quantitative (SA/Q) Skills, and four unit examinations were obtained. These four unit examinations covered approximately two-thirds of the first year program. Data from the final three units as well as the year-end comprehensive examination were not available at the time this preliminary study was conducted, but will be added to the present findings as soon as they become available.

RESULTS

Table 1 displays the intercorrelations (validity coefficients) between the subparts of the New MCAT and each of the four unit examinations administered to date. Descriptive statistics for both sets of examinations as well as internal consistency estimates (KR-20) of unit examinations scores are provided. The validity coefficients reported in this table have been corrected for the unreliability of the unit (criterion) examination scores in accordance with a formula found in Guilford and Fruchter (1978). The essence of this correction is to describe the correlations which would have been obtained had the unit examination been "perfectly reliable," that is, without measurement error.

Table 2 presents a breakdown of each unit examination into specific content areas (subtests). The first column (labeled %) represents the percentage of items within each unit examination allotted to the subject area. The validity coefficients reported in this table, as was the case in Table 1, have been corrected for the unreliability of scores on the subtests of the unit examination.

In each table the correlation coefficients are statistically significant, $p<.05$, unless otherwise noted (NS). Also, in Table 2, the highest correlation between subtest scores on the New MCAT and each subtest of the four unit examinations has been underscored.

A word of caution regarding the statistical significance of these validity coefficients is in order. With a sample size of just over 250 students, and employing a one-tailed test of significance, a correlation coefficient of approximately .10 would be statistically significant ($p<.05$). Since a correlation of .10 would imply that only 1% ($r^2=.01$) of the variation in the scores on the criterion measure (i.e. unit examinations; or its subtests) can be accounted for by variation in the scores on the subtests of the NEW MCAT (i.e. predictor measures), our attention should focus more on the practical, rather than the statistical, aspects of these correlations.

DISCUSSION

This paper represents the effort of one institution to examine the relationship between performance on the subparts of the New Medical College Admission Test and performance on examinations administered during the first year of medical study within an integrated curriculum. The results, however preliminary they may be, indicate at least a moderate degree of correlation between scores on the New MCAT and the four unit examinations administered at this institution, with this type of curriculum, and with the types of students enrolled here. Furthermore, the findings appear to be reasonable when viewed from a content perspective. For example, scores on the Chemistry (CHEM) subtest of the New MCAT show significant (both statistically and intuitively) correlations with such areas as Biochemistry and Renal and Respiratory structure and function.

Turning to other findings, performance in the areas of Biology knowledge and Science Problems (followed closely by performance on the Chemistry knowledge subpart of the New MCAT) were found to be consistently highest in terms of their correlations with performance on the unit examination subtests. Out of the 22 subtests on the four unit examinations, Biology knowledge (BIOL) correlated most highly with performance on eight of the subtests; followed by Science Problems (PROB), highest on seven out of the 22 subtests; and Chemistry knowledge (CHEM) highest on five out of the 22 subtests.

On the other hand, performance on the Physics knowledge (PHYS) and the Skills Analysis subparts (SA/R and SA/Q) of the New MCAT were observed to correlate highest less frequently with subjects of the unit examinations. For example, the Skills Analysis, Quantitative subtest (SA/Q) correlated highest with only two of the 22 unit exam subtests, Physics knowledge (PHYS) with only one, and Skills Analysis, Reading (SA/R) correlated highest with none of the 22 subtests. In fact, the Skills Analysis, Reading (SA/R) had the lowest correlation with performance in 17 out of the 22 subtests.

The higher correlations observed for Biology (BIOL) and Chemistry (CHEM) knowledge as well as the Science Problems (PROB) subpart of the New MCAT may well be explained by the overlapping items in the three areas. Since scores in each of the three science knowledge areas include items which are scored separately as Science Problems, such a finding may not be surprising.

CONCLUSIONS

This study investigated the relationship between scores on the New MCAT and performance on four unit examinations administered to first year medical students. The results of this preliminary investigation revealed statistically significant, positive correlations of moderate size. Furthermore, the pattern of intercorrelations, for the most part, revealed a tendency for higher correlations to be associated with a similarity between the major subject areas (e.g. Biology, Chemistry, or Physics), or type of skills required (e.g. problem solving or quantitative skills) on the unit examination and the corresponding part of the New MCAT. However, any interpretation of these findings must take into consideration several limitations including (1) the possibility of a restriction of range occurring on the New MCAT due to the selection process and (2) the particular nature of this institution, its curriculum, and students. Generalizing these findings to other institutions, curricula, and students would thus, be difficult. However, the generalizability of these findings may be enhanced by means of similar studies conducted at other institutions having different characteristics.

New MCAT scores are one of several predictive criteria utilized in the medical school admissions process. In order to obtain an accurate assessment of the contribution made by the New MCAT, the scores will have to be examined in relation to other predictive variables, such as undergraduate grade point average, demographic variables, and results of admissions interviews.

Since at least part of the rationale for the development of the New MCAT was a desire to upgrade the prediction of clinical performance, the correlation of New MCAT scores with measures of performance in the clinical and postgraduate years of training may be of particular interest. Investigations of this type, as well as further research into the prediction of performance in pre-clinical training, will be necessary to establish and clarify the practical utility of the New MCAT.

REFERENCES

Association of American Medical Colleges. New Medical College Admission Test Interpretative Manual. Washington, D.C.: AAMC, 1977.

Bartlett, J. W. Medical school and career performances of medical students with low medical college admission test scores. Journal of Medical Education, 1967, 42, 231-237.

Ceithaml, J., The Admissions Process, in Purcell, E. F. (Ed.), Recent Trends in Medical Education. New York: Josiah Macy, Jr. Foundation, 1976.

Erdmann, J. B. Separating wheat from chaff: Revision of the MCAT. Journal of Medical Education, 1972, 47, 747-749.

Fredericks, M. A., & Mundy, P. The relationship between social class, average grade in college, Medical College Admission Test scores, and academic achievement of students in a medical school. Journal of Medical Education, 1967, 42, 126-133.

Garfield, S. L., & Wolpin, M. MCAT scores and continuation in medical school. Journal of Medical Education, 1961, 36, 888-891.

Gough, H. G., Hall, W. B., & Harris, R. E. Admissions procedures as forecasters of performance in medical training. Journal of Medical Education, 1963, 38, 983-998.

Guilford, J. P., & Fruchter, B. Fundamental Statistics in Psychology and Education (6th ed.). New York: McGraw-Hill, 1978.

Haley, H. B., Juan, I. R., & Paiva, R. E. A. MCAT scores in relation to personality measures and biographical variables. Journal of Medical Education, 1977, 26, 947-958.

Richards, J. M., Review of the MCAT, in Buros, O. K. (Ed.). The Seventh Mental Measurements Yearbook. Highland Park, N.J.: Gryphon Press, 1972.

Watson, R. I. Predicting academic success through achievement and aptitude tests. Journal of Medical Education, 1955, 30, 383-390.

Table 1

Descriptive Statistics and Validity Coefficients[1] between New MCAT Subtests and Four Unit Examinations

	INTRO Unit	CV Unit	GI Unit	RESP/RENAL Unit	MEAN	STD DEV
BIOL	31	38	33	31	9.84	1.56
CHEM	25	40	29	39	9.98	1.76
PHYS	15	33	23	27	9.63	2.19
PROB	24	42	32	36	9.82	1.73
SA/R	08^{NS}	20	15	17	9.05	1.71
SA/Q	19	33	29	27	9.48	1.81
N of STUDENTS[2]	250	251	251	251		
N of ITEMS	328	189	147	115		
MEAN	241.29	147.44	117.54	81.76		
% MEAN	73.56	78.01	79.96	71.10		
STD DEV	22.42	13.44	10.90	11.00		
RELIABILITY (KR-20)	.91	.86	.83	.85		

[1] Unless otherwise noted, all validity coefficients are statistically significant with $p<.05$. Decimal points have been omitted.

[2] Two students were absent from the Unit 1 exam; one student was absent in Units 2 through 4. r's are based on these N's

NS: statistically non-significant: $p>.05$

Table 2

Validity Coefficients[1] between New MCAT Subtests and Subtests of Four Unit Examinations[2]

EXAMINATION/SUBTEST	%	BIOL	CHEM	PHYS	PROB	SA/R	SA/Q
INTRO UNIT							
Histology Practical	30	40	31	18_{NS}	29	07^{NS}	18
Biochemistry - I	22	23	17	06^{NS}	19	16_{NS}	10
Biochemistry - II		26	30	20_{NS}	27	08^{NS}	16
Histology	15	29	21	09^{NS}	21	17	25
Physiology	03	28	32_{NS}	36_{NS}	32_{NS}	19_{NS}	34
Gross Anatomy	30	15	08^{NS}	04^{NS}	08^{NS}	-02^{NS}	10
CV UNIT							
Muscle, CV	08	42	48	36	43	19	25
Cardiology	21	42	43	43	46	16	33
Circulation	09	29	52	35	44	25	44
Hemostasis	06	12	11	10	15	14	10
Biochemistry	16	33	29	17	28	17_{NS}	17
Anatomy	10	31	28	21	32	08^{NS}	24
Neurophysiology	03	70	70	57	79	19	76
Gross Anatomy	26	26	24	23	31	22	35
GI UNIT							
Histology	20	29	22	14	29	14	25
Nutrition	08	25	25	25	19	28	38
Biochemistry	14	46	33	36	40	17_{NS}	28
Physiology	12	34	29	21	25	-00^{NS}	24
Gross Anatomy	13	28	30	22	35	16	25
Practical	33	15	15	14	24	11	14
RESP/RENAL UNIT							
Renal	43	24	38	28	36	16	30
Respiratory	57	34	38	25	34	18	24

---

[1] Unless otherwise noted, all validity coefficients are statistically significant with $p<.05$. Decimal points have been omitted.

[2] Largest row correlations are underscored

A FACTOR ANALYTIC STUDY OF THE OLD AND NEW MCAT EXAMINATIONS

John B. Molidor, Ph.D. and Arthur S. Elstein, Ph.D.

Michigan State University[1]

Introduction

Since the development of the Medical College Admissions Test (MCAT), scores on this examination and grade point averages (GPA) have been the principal quantitative measures for selecting students for medical schools from the large applicant pool. The MCAT was developed as a means of verifying the grade record by providing a mechanism by which medical schools could rate applicants on a standardized instrument, could correct for variation in the quality of undergraduate institutions, and could adjust for differences in grading standards and in the level of difficulty of undergraduate courses. It has also been used to provide estimates of overall ability of students whose early college performance was not particularly distinguished but who showed substantial gains in academic achievement in the junior and senior years--the so-called "late bloomers."

In the 60's and 70's, dissatisfaction with the test grew (Bartlett, 1967; Gough, Hall and Harris, 1963; Turner, Helper and Krisha, 1974; Wingard and Williamson, 1973). These critics argued that while the test was successful in predicting academic performance in the first two years of medical school, it was not particularly successful in predicting clinical performance in the last two years of medical school or in subsequent practice. They also claimed that the test overemphasized recall of specific facts, did not sufficiently stress analytic skills and comprehension of scientific principles, and did not measure such personal qualities relevant to the practice of medicine as empathy, responsibility and flexibility.

In response to these criticisms, the Association of American Medical Colleges (AAMC) sponsored the development of a new version of the MCAT. The new MCAT (AAMC, 1977) was first used extensively in processing applications for medical school classes that matriculated in September, 1978. Since the old MCAT had been in use for over twenty-five years, it was well known to all medical school admissions officers. The new test differs in three obvious respects: It yields six subtest scores instead of four, the scores are reported on a fifteen-point scale with a mean of eight, and it takes about twice as long to administer. But does the new test differ in more fundamental, less obvious ways?

---

[1] We wish to thank Judith W. Krupka, Ph.D., for her assistance and cooperation in related studies and in the preliminary phases of this research, and Verda Scheifley, Ph.D., for consultation on the statistical analysis. Reprint requests to: John B. Molidor, Ph.D., College of Human Medicine, Office of Admissions, A-232 Life Sciences, Michigan State University, East Lansing, Michigan 48824.

Factor analysis is a multivariate statistical technique widely used to gain insight into the differences and similarities between two tests that on the surface appear to measure overlapping domains (for instance, tests of intelligence). Faced with the problem of trying to understand better the scores yielded by the new MCAT, we decided to compare the old and new examinations by means of factor analysis.

This paper reports our analysis of the factorial structure of the new MCAT, compares that structure with the factorial structure of the now discontinued old MCAT, and considers how the new test might best be used as a selection device in medical school admissions.

Method

Subjects

Data were available on 2880 students who took the old MCAT at either of the two 1976 test administrations and who applied to the class entering in September, 1977, at the College of Human Medicine at Michigan State University. For the class entering in Fall, 1978, who took the new MCAT, data were collected on 2391 students examined at either of two 1977 test administrations. Some students took the examination twice in an effort to improve their score. In these cases, only the more recent administration was used, so each student is represented only once.

Procedures

The old MCAT yielded four scores on each subject: Verbal, Quantitative, General Information, and Science. The new MCAT yields six scores for each examinee. Three of these scores are in the area of Science Knowledge, including assessments in Biology, Chemistry, and Physics. The fourth score is Science Problems, an assessment of knowledge and problem solving ability in Biology, Chemistry, and Physics. Reading Skills and Quantitative Skills are the fifth and sixth areas assessed by this new examination. Detailed descriptions of these subtests are provided in the manual for the new MCAT (AAMC, 1977). Note that there is item overlap and non-independence between the Science Knowledge subtests and the Science Problems subtest. The extent of the item overlap is unknown to us. However, the MCAT manual states that the intercorrelations demonstrate that the abilities measured by the various areas of assessment are sufficiently different to justify their inclusion and the reporting of six discrete scores.

To analyze the old MCAT scores, the four subtests of the old MCAT were intercorrelated and the resulting 4 X 4 correlation matrix was factor analyzed using the SPSS program employing principal components analysis (Nie et al., 1975). The solution was then transformed using a varimax rotation.

The six subtests of the new MCAT were intercorrelated and the resulting 6 X 6 correlation matrix was also factor analyzed using the same program.

Results

The results of the factor analysis of the old MCAT scores are presented in Table 1. The results display a two-factor varimax solution. The first factor has high loadings on the Verbal and General Information subtests. This factor appears to be assessing a verbal ability dimension. The second factor has high loadings on the Quantitative and Science subtests of the old MCAT. This factor appears to assess the science-quantitative content domain. The first factor accounts for 71.5% of the total variance (eigenvalue = 2.86), and the second factor accounts for an additional 16.8% of the total variance (eigenvalue = .67).

The results of the factor analysis of the 6 X 6 matrix of new MCAT scores, displayed in Table 2, also yield a two-factor solution. However, in our judgment, the solution is not as well-defined as the two-factor solution of the old MCAT. The first factor has high loadings on the Biology, Chemistry, Physics, and Science Problems subtests. This factor appears to be measuring a science knowledge dimension. The second factor has high loadings on the Reading and Quantitative Skills subtests. This factor appears to assess a skills analysis domain. The first factor accounts for 73.4% of the total variance (eigenvalue = 4.40), and the second factor accounts for an additional 9.8% (eigenvalue = .59).

Discussion

Discussion of the results of these analyses centers on two major questions:

1. What do we know about the content domains assessed by these tests?
2. How can one best use the new MCAT as a selection device?

Content Domains

The results of the factor analyses indicate that the old and new MCAT are two-factor tests. In the old MCAT, the Verbal and General Information subtests accounted for the greatest amount of total variance. In the new MCAT, the Science Knowledge and Science Problems subtests account for the largest portion of variance. Thus, we find the new MCAT is heavily saturated with scientific-quantitative content.

Although the old MCAT had scientific-quantitative content similar to the new exam, it also measured educational achievement and ability in the verbal domain. With the possible exception of the Skills Analysis Reading subtest, the new MCAT examination largely ignores assessment of verbal abilities. The Reading subtest of the new MCAT may prove to be a useful contribution to selecting applicants, although its correlations with other college-level reading tests are not yet published.

It is regrettable that the new test samples broad education and acquaintance with culture and literature less effectively than the old MCAT. The old MCAT was criticized for over-emphasizing recall of specific facts and not sufficiently stressing analytic skills and comprehension of scientific principles. The new MCAT may be criticized as essentially a test of knowledge and application

Table 1

FACTOR ANALYSIS OF OLD MCAT SCORES

(n = 2880)

Old MCAT Subtests	Factor I	Factor II	h^2
Verbal	81	35	78
Quantitative	28	80	72
General Information	81	32	76
Science	42	73	79
Total Variance Accounted For	71.5%	16.8%	

Table 2

FACTOR ANALYSIS OF NEW MCAT SCORES

(n = 2391)

New MCAT Subtests	Factor I	Factor II	h^2
Science Knowledge - Biology	69	44	68
Science Knowledge - Chemistry	78	42	80
Science Knowledge - Physics	73	44	73
Science Problems	85	43	89
Skills Analysis - Reading	35	66	56
Skills Analysis - Quantitative	45	76	79
Total Variance Accounted For	73.4%	9.8%	

(Decimal points are ommitted)

of biological and physical science facts and principles and of quantitative skills. It slights the importance of social science and humanities as preparation for a medical career.

The message of the new test seems to be unmistakable: If a student wishes to enter a medical school, achievements in biology, chemistry, physics, and quantitative skills are more important than ever. The lack of emphasis upon comparable academic achievement in the social sciences and humanities is striking, especially in view of repeated calls for broadening premedical education beyond the biological sciences (Thomas, 1978). The underlying conception of the "complete physician" is narrow and dominated by the same technical outlook the profession is now trying to balance in many new programs.

Use as a Selection Device

A very practical consideration for admissions officers is how to use the new MCAT as a selection device. Based on the factor analysis of the new MCAT, we recommend that two mean scores be calculated, corresponding to the two factors identified. The first score is an average of the four Science Knowledge and Science Problems subtests and serves as an assessment of science knowledge. The second score is an average of the Reading and Quantitative Skills subtests, and represents an assessment of analytic skills. The use of two scores drastically reduces the possibility of information overload without diminishing the power of the test to predict medical school achievement.

Once these two means have been calculated, they can be weighted differentially to reflect varying committee policies. For example, the science knowledge mean score could be given twice the weight the skills analysis mean score. However, we recommend that the two scores be weighted equally for two reasons. First, equal weights reflect approximately the amount of time spent on each set of subtests; 170 minutes are sepnt on the Science Knowledge and Science Problems subtests, and 180 minutes are allotted for the Reading and Quantitative Skills subtests. Since nearly equal time is spent on each section, it seems reasonable that equal weight be given to each section.

The second reason for giving equal weight to the two mean scores is that it allows an assessment of another area besides the sciences, namely analytic skills. If all six subtests of the new MCAT were added without weighting to construct a composite score, the sciences would be weighted too heavily, judging by the relative variance accounted for by the two factors. Using two mean scores and assigning equal weight to each allows a broader assessment of individual applicant's abilities.

References

Association of American Medical Colleges, Division of Educational Measurement and Research. *New Medical College Admission Test: Interpretive Manual.* Washington: Author, 1977.

Bartlett, J.W. Medical school and career performances of medical students with low Medical College Admission Test scores. *Journal of Medical Education*, 1967, *42*, 231-237.

Gough, H.G., Hall, W.B., & Harris, R.E. Admissions procedures as forecasters of performance in medical training. *Journal of Medical Education*, 1963, *38*, 983-998.

Nie, N., Hull, C., Jenkins, J., Steinbrenner, K., & Bent, D. *Statistical Package for the Social Sciences* (2nd Edition). New York: McGraw-Hill, 1975.

Thomas, L. Notes of a biology watcher: How to fix the premedical curriculum. *New England Journal of Medicine*, 1978, *298*, 1180-1181.

Turner, E.V., Helper, M.M., & Krisha, S.D. Predictors of clinical performance. *Journal of Medical Education*, 1974, *49*, 338-342.

Wingard, J.R. & Williamson, J.W. Grades as predictors of physicians' career performance: An evaluative literature review. *Journal of Medical Education*, 1973, *48*, 311-322.

A Non-Linear Discriminant Analytic Approach to Predicting Medical Student Performance*

Charles P. Friedman
Claudia M. Cheatham
Carol Q. Porter
William E. Bakewell

University of North Carolina at Chapel Hill

Due largely to the advent of AMCAS and similar computerized systems, medical school admissions committees have at their disposal vast quantities of objective data about applicants. This plethora can be both a blessing and a curse. Computerization makes the processing of information faster and more economical; but the amount of information to work with can be overwhelming to committee staff and members. For this reason, numerous methods of reducing the available data to one or more predictive indices have been reported.[1] Such methods have been validated by using the index either as a predictor of actual performance of medical students through longitudinal studies,[2,3,4,5] or as a predictor of what admissions committees would decide were these committees given the data in unreduced form.[6,7,8] Such research has shown that reduced, quantitative admissions data can predict both the actions of admissions committees and the performance of medical students, especially in the early curricular years.

This study presents a related but new methodology for predicting medical student performance from admission data.** The effort was motivated by two characteristics of existing methods reported in the literature. First, most prediction studies employ a regression procedure that assumes linear relationships between predictor and performance variables. Yet, there exist few theoretical justifications to restrict relationships between admissions variables to a strictly linear form, particularly when some of the variables involved are not measures of academic aptitude.[9] Thus, a methodology which incorporates non-linearity can potentially yield greater predictive power. Second, most studies reported in the literature predict grades, class rank, or National Board examination scores, despite a prevalent notion that such measures do not, by themselves, indicate success in medical school or thereafter.

The method described here operationalizes medical school achievement categorically, similar to the work of Bartlett.[10] Educational outcomes for each student are classified as (1) unsatisfactory, (2) satisfactory, or (3) honors. At the University of North Carolina, local and nationally produced tests, reports of clinical performance, personal letters, and other sources contribute to committee decisions that distinguish between students in each of these groups. Thus, the categorical performance outcome is seen as an indicator of achievement having greater validity than any of its contributory sources. Because the categorization of each student is predicated on performance in the entire medical curriculum, it reflects both basic science and clinical achievement. Also, it is consistent

---

* Reprints avaliable from: The Office of Medical Studies, University of North Carolina School of Medicine, Chapel Hill, North Carolina 27514

** This method is not currently in use by the admissions committee at the University of North Carolina. Reported here is a developmental study that may lead to a decision to adopt the method or some variation of it.

164

with the approach to medical student evaluation at UNC where students receive one summary grade for each year of the curriculum.

Based on admission and performance data of past students, the method can compute, for each prospective student, the probability of ultimate membership in each of the three outcome groups. Discriminant function analysis forms the core of the prediction algorithm,[11] with non-linear effects of predictor variables systematically incorporated. While discriminant function analysis has been employed in several studies which validate data reduction procedures by predicting admission committee decisions,[6,8] it has not, to the authors' knowledge, been employed in predictive studies based on categories of student performance in medical school.

Method

Data for the predictive study are derived from the classes entering the University of North Carolina School of Medicine from 1973 through 1975 (N=398). The predictor variables shown in Table 1 were taken from the AMCAS tapes and from a biographical questionnaire administered to matriculating medical students. The authors constructed two interaction variables hypothesized to be important predictors of performance: (a) the product of total GPA and selectivity of the student's undergraduate school, and (b) the product of total undergraduate science hours and science GPA. It was further hypothesized that each interaction term would be a more powerful predictor than the variables from which it was constructed. Also, the average annual change of each student's science and nonscience GPA was computed, on the assumption that students whose GPA's improve or decline over time may perform differently from those exhibiting stable performance. Each student was assigned to one of three outcome categories based on knowledge of later performance, as follows:

> Group 1 (Unsatisfactory): Students who either repeated a year of the curriculum, were dismissed from school, were assigned to a special remedial program, or were asked to retake one or more subtests of the National Board Part II examination
>
> Group 2 (Satisfactory): Made standard progress through the curriculum but were not elected to the medical student honor society (AOA).
>
> Group 3 (Honors): Made standard progress through the curriculum and were elected to AOA

Discriminant fuction analysis was employed to determine the degree to which categorized student performance could be accounted for statistically by the predictor variables. A stepwise procedure was used to determine the relative power of the predictor variables and also to generate the most parsimonious list of predictors to facilitate economical computation in the future.

Prior to execution of the discriminant function analysis, each predictor with an interval metric was tested for polynomial trends. This was accomplished by regressing each such predictor, its square and its cube on a quasi-interval variable coded "1" for Group 1, "2" for Group 2 and "3" for Group 3. This procedure indicated that a better fit to the performance data could be achieved by employing higher-order polynomial terms for five of the predictor variables, as shown in Table 1. In these cases, a polynomial function of the linear variable (using unstandardized coefficients generated by the regression discussed above) was substituted for the linear variable in the discriminant analysis, on the assumption that this would enhance the statistical power of the procedure.*

Results

The step-wise analysis procedure resulted in construction of two highly significant discriminant functions ($p < .005$) comprising 11 predictor variables as shown in Table 2. The first variable entered in the step-wise procedure was the interaction term of GPA weighted by selectivity of undergraduate school. The science MCAT, entered second, emerged as the most important predictor among the old MCAT subtests, as it has in other studies.[12] Four of the variables entered were in fact the best-fit polynomials computed through the earlier series of regression procedures.

Some care is required to interpret the effects of the various predictor variables because of the numerous non-linearities. For example, the square of the change in non-science GPA, entered third in the analysis, contributed positively and most heavily to the second discriminant function. Thus, either an increase or decrease over time in non-science GPA tended to raise the value of the second function; however, higher values of the second function were associated with membership in Group 2 (satisfactory). The net effect of this predictor variable is for extreme high and low values to increase membership probability in Group 2.

Based on the discriminant function, probability of membership in each of the groups is assigned for each subject. If one assumes that each subject is "predicted to be" in the group with highest probability of membership, the two discriminant functions were able to "classify" correctly 79.1% of cases, when a Bayesian adjustment in group membership probabilities is made by setting "prior" probabilities proportional to group sizes. The procedure was most efficient in predicting satisfactory performance: 288 of 351 students (82%) predicted to be in Group 2 actually were. It was less efficient in predicting unsatisfactory performance: 20 of 33 students (61%) predicted into Group 1 actually were. Honors students could not be distinguished from "satisfactory" students.

When the same discriminant analytic procedure was followed without the polynomials, only one significant discriminant function was derived. This is understood by observing that many of the variables loading heavily on the second function have strong non-linear relations to the group membership outcome variable. Omission of the higher-order polynomials from the analysis did not diminish the number of cases correctly classified by the discriminant function, however. This is explained by examination of a plot of cases in discriminant function space, which reveals that most of the cases not correctly classified are significant outliers unaffected by the information contained in the second discriminant function.

---

*This remained an assumption because of the unconventional treatment of the outcome variable in performing the regressions.

Discussion

The discriminant analytic procedure makes it possible for admissions committees to assign each applicant a probability of membership in Group 1 (unsatisfactory), Group 2 (satisfactory), and Group 3 (honors) based upon data available about the student and the aggregate behavior of past students in the school taking essentially the same curriculum. For example, one subject in the study was assigned a .249 probability of being in Group 1, a probability of .740 of being in Group 2, and a .011 probability of membership in Group 3. For classification purposes, this subject was "predicted" to be in Group 2. But this also tells the committee that, based on performance of past students, the sample subject has a one-in-four chance of having academic problems when all significant predictor variables (Table 2) are taken into account. The authors assert that this format for quantitative data about applicants is potentially more useful than predicted GPA or predicted National Board scores. Such a procedure retains the advantages of standard prediction methods in that it reduces the plethora of data to a manageable size and makes the relative predictive power of each variable an empirical question resolvable by examining behavior of past students.

While the analysis performed here of necessity employs the old MCAT, there is ample reason to believe that the new MCAT will retain predictive power.[13] The authors are investigating the power of the new MCAT to identify students who will have academic trouble in the early years of medical school. As data accumulates, the validity of the new MCAT for this purpose will become clearer.

The fact that five of the predictor variables display significant non-linear relationships to the outcome measures suggests that non-linearity be incorporated into prediction algorithms. It also suggests that admission committees look carefully at their assumptions about what is "good" in a student's application. Variables displaying significant polynomial terms to the second and third order are not monotonic predictors over their entire functional domain, and thus it is not always "better" to have a higher value of certain predictors.

The authors do not envision this parsimonious, automated approach to analysis of application data as an end in itself leading automatically to admission decisions that maximize discriminant functions. Rather it is seen as a means to a different end: to help admissions committees devote more of their increasingly scarce time and resources to in-depth examination of the "softer"--and potentially more important--aspects of each individual's application. Jason cautions us against

> "selection... primarily directed at finding individuals who merely are likely to survive the first year of medical school, where virtually all of the attrition ordinarily occurs but where few of the characteristics of the effective physician are required for success."[14]

Knowing the projected probability of overall academic success for each student, admission committees can focus their efforts on a broader search for the effective physician. When committees choose to admit academically "risky" students, they can do so with a quantitative estimate of the amount of risk they are taking.

References

1. Cuca, Janet L., et al. *The Medical School Admissions Process: A Review of the literature 1955-1967*. Special report published by the Division of Student Studies, Association of American Medical Colleges, October, 1976.

2. Best, William R., et al. Multivariate Predictors In Selecting Medical Students. *J. Med. Educ*., 46:42-50, 1971.

3. McGuire, Frederick L. Fifteen Years of Predicting Medical School Performance. *J. Med. Educ*., 52:416-417, 1977.

4. Gough, Harrison G., et al. Admissions Procedures as Forecasters of Performance in Medical Training. *J. Med. Educ.*, 38:983-998, 1963.

5. Jorve, Susan. *The Relationship Between Objective Pre-Admission Measures and Academic Achievement in the Basic Sciences*. Unpublished manuscript, Univ. of North Dakota, 1978.

6. Cullen, Thomas J. *The Use of The Discriminant Function as a First Screen of Applicants to Medical School*. Paper presented at the 13th RIME Conference, Chicago, 1974.

7. Weisman, Robert A., et al. On Achieving Greater Uniformity in Admissions Committee Decisions. *J. Med. Educ*., 47:593-602, 1972.

8. Milstein, Robert M. et al. Prediction of Screening Decisions in a Medical School Admission Process. *J. Med. Educ*., 51:621-633, 1976.

9. Guilford, J. P. "Validations of Measures of Interest and Temperament." In L. L. Thurstone (Ed.), *Applications of Psychology*, New York: Harper, 1952.

10. Bartlett, J. W. Medical School and Career Performances of Students with Low MCAT Scores. *J. Med. Educ*., 42:231-237, 1967.

11. Van de Geer, John P. *Introduction To Multivariate Analysis For The Social Sciences*. San Francisco: W. H. Freeman and Company, 1971.

12. Stefanu, Constantine, and Farmer, Albert T., Jr. The Differential Predictive Validity Of Science MCAT in the Admission Process. *J. Med. Educ*., 46:461-463, 1971.

13. *New Medical College Admission Test Interpretive Manual*. Washington, D.C.: AAMC, 1977.

14. Jason, Hilliard. Editorial: The Admissions Process in Medicine. *J. Med. Educ*., 47:663, 1972.

Table 1. Predictor Variables

Personal

- Sex (1: female, 2: male)
- Age[1]
- Father's Education[1]
- Mother's Education[1]
- Marital Status (1: married, 2: single)
- Size of Home Community

Academic

- Science & Non-science GPA
- Four "old" MCAT subtests
- Selectivity of Undergraduate School[1,2]
- Total Undergraduate Science Hours
- "Quantitativeness" of undergraduate major[3]

Interaction & Non-Linear Terms

- (Selectivity) X (Total GPA)
- (Total Science Hours) X (Science GPA)
- Average year-to-year change:
  - in science GPA[4]
  - in non-science GPA[1,4]

[1] Indicates a variable with significant higher-order polynomial fit to outcome data.

[2] Computed from the average total SAT of each school's entering class.

[3] Coded on a 6 point scale with mathematics and physics "most quantitative" and the humanities "least quantitative."

[4] Computed by averaging the change from year 1 to year 2 with the change from year 2 to year 3.

Table 2. The Discriminant Functions

Order of Entry	Variable	Standardized Coefficients* Function 1	Function 2
1	Selectivity X GPA	-.05	-2.01
2	Science MCAT	.34	-.24
3	Average change in non-science GPA§	-.14	.37
4	Father's education†	.16	.51
5	Non-science GPA	.28	1.04
6	Selectivity†	.33	1.60
7	Science GPA	.32	.38
8	Age†	.01	.45
9	GI MCAT	.13	-.39
10	Sex	-.15	.07
11	Marital Status	-.12	.08

*Higher values of Function 1 are associated with higher levels of performance; however, higher values of Function 2 are associated with higher probability of membership in Group 2.

§Quadratic polynomial function of this variable used to generate discriminant functions.

†Cubic polynomial function of this variable used to generate discriminant functions.

A STRUCTURED INTERVIEW FOR THE SELECTION OF PHYSICIAN'S ASSISTANT STUDENTS

Bruce R. Niebuhr, Ph.D., Steven R. Shelton, P.A.-C.,
and Shirley A. Werner, P.A.-C.
The University of Texas Medical Branch at Galveston

The use of the personal interview as a selection device by medical educational programs has been sharply criticized as lacking reliability and validity (1-4). One residency program abandoned use of the selection interview upon finding that agreement between interviewers in rating applicants was virtually nil (4). Some medical schools have also eliminated the interview (5). Nevertheless, the selection interview continues to be extensively used, in part because it serves other purposes, such as a public relations function for the educational programs (6). In any case, because of its widespread use Morse maintains that more research is needed to determine effective ways to produce reliable data from the selection interview (2). In one recent study significant inter-rater agreement was obtained from interviews of Physician's Assistant (P.A.) applicants (7).

In general, the results of published studies do not support the validity of selection interviews in medical programs (2,3,8,9). An exception is a study by Murden et al in which interview ratings, but not premedical GPA, were found to correlate with later clinical success in medical school (10).

In the Fall of 1978 the faculty of the Physician's Assistant Program of The School of Allied Health Sciences, The University of Texas Medical Branch at Galveston, began a reevaluation of the student selection process. In order to rank applicants a selection formula was developed which included as factors the selection interview, the Otis Mental Ability Test (11), the Nelson-Denny Reading Test (12), and three variants of college grade point average (GPA): overall GPA, GPA in scientific subjects, and GPA for the last 30 semester hours. Of particular concern to the faculty was the selection interview, viewed as a method to examine non-cognitive characteristics of the applicants not assessed by grade point average and standardized aptitude tests. The consensus of the faculty was to retain a selection interview and to structure it to improve its reliability. The faculty were requested to identify question categories they considered important to an interview. The categories ultimately agreed upon to form the basis of the structured interview are presented in Table 1. The categories were selected with an awareness on the faculty's part that interviewers would have differing interpretations of the applicant's responses to questions. Whether the interviewers' interpretations would be similar enough to yield reliable interview ratings is an empirical question.

The purposes of the present paper are to describe the structured interview format, to assess the reliability of interview ratings, and to make a preliminary examination of the validity of the ratings. In particular, concurrent validity was examined by comparing interview ratings with scores on the Hogan Empathy Scale (13). Empathy is one personality characteristic

---

Reprint requests may be sent to Dr. Bruce R. Niebuhr, Office of Academic Affairs, School of Allied Health Sciences, The University of Texas Medical Branch at Galveston, Galveston, Texas 77550.

170

described as playing an important role in effective health care (14).

Method

A total of 94 applicants met or expected to meet the minimum application standards of 60 semester hours in specified courses with a minimum GPA of 2.0 on a 4-point scale and were invited to Galveston for testing and interviewing. Seven of these applicants had applied in previous years but had not been accepted. The demographic characteristics of the applicants presented in Table 2 are similar to those reported for other programs (15). The size of the incoming class was set at 25.

Interviews were conducted by 17 full and adjunct faculty members of the Physician's Assistant program. Of the faculty, eight were P.A.s, three were physicians, and four were Ph.D.s. Two program graduates, not employed at the University, also served as interviewers. The interviewers were provided with a guide to the interview which specified the question categories (Table 1) and provided space for applicant responses and for additional questions and responses. In addition, the interview guide included a checklist for the interviewers to assess the following behavioral characteristics of the applicants: dress, mannerisms, eye contact, speech fluency, and responsiveness to interviewer's questions. A brief summary of each applicant was attached to the guide which included GPA and colleges attended. In order to facilitate a free flow of conversation, the interviewers were not required to ask questions in any predetermined order.

Each interview was allotted one-half hour and was conducted by one interviewer. Upon completion of the interview the interviewer rated the applicant according to the 5-point scale presented in Table 3. Each applicant was interviewed three times by different interviewers. Since each interviewer rated the applicant before discussing the interview with others the interviews were performed independently. In order to minimize the effect of any tendency on the part of any group of interviewers to rate consistently high or low, assignments were made in such a manner that a group of any three interviewers would not interview more than one applicant. A mean rating of the three interviews was calculated, and included as a factor in the selection formula.

In a testing session the applicants were administered the Otis Mental Ability and Nelson-Denny Tests. In addition, they were asked to complete the 64-item Hogan Empathy questionnaire (13). The empathy questionnaire was administered as part of an ongoing research study on personality characteristics of P.A. students and was not used as a selection factor.

Because of the limited number of interviews common to any two interviewers, traditional correlation type inter-rater reliability coefficients were not utilized. Reliability was assessed utilizing a chi-square goodness-of-fit test in the following manner. For each set of three interviews the absolute number of points each rating deviated from the other two was obtained and summed. This deviation sum can take on a limited number of values: 0, 2, 4, 6, or 8. A deviation sum of 0 represents complete agreement among the three interviews. The worst-case sum of 8 can only occur if one rating is 5 and a second is 1 (the rating of the third interview will not affect the deviation sum in this case, e.g., 5, 1, 1, and 5, 1, 5 both yield deviation sums of 8). The chi-square goodness-of-fit test was used to determine whether across all interviews the ratings tended to show greater agreement than expected by chance. A similar use of chi-square to assess inter-rater reliability is described by Pharris and Van Cleve (7).

Concurrent validity of the interview was examined by obtaining Pearson product-moment correlations of interview ratings and Hogan Empathy Scale scores. Additional inter-correlations of ratings, GPA, Otis IQ, combined Nelson-Denny vocabulary and comprehension scores, and Empathy Scale scores were obtained.

Results

Of the 94 individuals invited, 87 were interviewed and tested. The remaining 7 either withdrew their applications or failed to appear for interviewing. Completed empathy questionnaires were obtained for 82 applicants. In the case of one applicant, an interview rating of "1" by one interviewer and "5" by another was obtained. In this case a fourth interview was conducted and a mean rating based upon the four interviews determined.

The mean interview rating for the 87 applicants was 3.6 with a standard deviation of .8. The results of the chi-square goodness-of-fit test to determine inter-rater reliability are presented in Table 4. The amount of agreement is significantly greater than that expected by chance ($x^2 = 145$, $df = 4$, $p < .001$). The interview ratings were found to correlate significantly with the Empathy Scale ($r = .25$, $df = 80$, $p < .05$). Inter-correlations are given in Table 5.

Discussion

Results of the present study support the reliability and concurrent validity of the selection interview for the incoming Physician's Assistant class of 1979-81 at the University of Texas Medical Branch. The predictive validity of the interview will be examined throughout the two-year curriculum. Of particular interest is whether the interview will be better than GPA, Otis, or Nelson-Denny as a predictor of clinical performance. Gough & Hall found that clinical performance of medical students was not predictable from entering GPA or aptitude scores (14).

Both the interview and the Otis were found to correlate significantly with the Hogan Empathy Scale. However, the interview, unlike the Otis Test, can be structured to more expressly examine empathic capabilities of applicants. The Hogan Empathy Scale itself is not feasible as a selection instrument. Unlike a standardized aptitude test, it is available in only one form, the scoring key for which was published in the original article (13). Personality inventories such as the Hogan Empathy Scale are most valuable as criteria for assessing the concurrent validity of a selection interview.

References

1. Kelly E. L. A critique of the interview. J. Med. Educ., 1957, 32, 78-84.
2. Morse P. K. The interview as a means of assessing personality characteristics. Proc. Conf. Pers. Meas. Med. Educ., 1971, 51-63.

3. Rimm A. A., Pazdral M. A., Sine B. S. Methodological study of rating applicant interviews. *J. Med. Educ.*, 1968, 43, 1085.
4. Gordon M. J., Lincoln J. A. Family practice resident selection: Value of the interview. *J. Fam. Prac.*, 1976, 3, 175-177.
5. Char W. F., McCermott J. F., Haning W. F., Hansen M. J. Interviewing, motivation, and clinical judgment. *J. Med. Educ.*, 1975, 50, 192-194.
6. Ulrich L., Trumbo, L. The selection interview since 1949. *Psyc. Bull.*, 1965, 63, 100-116.
7. Pharris J. L., Van Cleve H. P. The personal interview for applicant selection. *J. Fam. Prac.*, 1977, 4, 377-378.
8. Gough H. G., Hall W. B. An attempt to predict graduation from medical school. *J. Med. Educ.*, 1975, 50, 940-950.
9. Richards J. M., Taylor C. W. Predicting academic achievement in a college of medicine from grades, test scores, interviews, and ratings. *Educ. Psyc. Meas.*, 1161, 21, 987-994.
10. Murden R., Galloway G. M., Reid J. C., Colwill J. M. Academic and personal characteristics as predictors of clinical success in medical school. *Proc. 16th Ann. Conf. Res. Med. Educ.*, 1977, 181-186.
11. Otis A. S. *Otis Quick-Scoring Mental Ability Tests*. New York: Harcourt Brace Jovanovich, 1965.
12. Nelson M. J., Denny E. C. *The Nelson-Denny Reading Test*. Boston: Houghton Mifflin, 1960.
13. Hogan R. Development of an empathy scale. *J. Cons. Clin. Psyc.*, 1969, 33, 307-316.
14. Anthony W. A., Carkhuff R. R. *The Art of Health Care*. Amherst, M. A.: Human Resources Press, 1976.
15. Crovitz E. Comparison of male and female Physician's Associate program applicants. *J. Med. Educ.*, 1975, 50, 672-676.
16. Gough H. G., Hall W. B. The prediction of academic and clinical performance in medical school. *Res. Higher Educ.*, 1975, 3, 301-314.

Table 1
Question Categories of the Selection Interview

1. How the applicant became interested in becoming a P.A.
2. How the applicant's career goals could be satisfied by being a P.A.
3. The most and least attractive features of being a P.A. as perceived by the applicant.
4. Previous experience in health care.
5. Use of P.A. education as a method to enter another field.
6. Had the applicant considered medical school.
7. Had the applicant applied to other P.A. programs.
8. Financial plans for educational costs.
9. Family support for becoming a P.A.
10. Applicant's feelings about the dependent role of the P.A. upon a physician.
11. Applicant's response to a hypothetical situation in which he or she was not given sufficient responsibility.
12. Interests outside of work and school.
13. Expected changes in lifestyle if admitted to the program.
14. Expected salary as a P.A.

Table 2

Interview Rating Scale

Rating	Description
1	Unacceptable. I do not want this student in the class.
2	Marginal. I am doubtful about this student.
3	Satisfactory. This student, while not outstanding, should be fine.
4	Very good. I am very enthused about this student.
5	Excellent. A magnificent prospect.

Table 3

Demographic Characteristics of Applicant Pool

Sex	Number	Age (in years)	
Male	48	Mean	25.0
Female	46	Standard Deviation	5.4

Previous Education	Number
Ph. D.	1
Masters	3
Baccalaureate	27
Associate	2
College Hours without Degree	60

Table 4

Chi-Square Goodness-of-Fit Test for Determining Inter-rater Reliability (Interviews of 87 Applicants)

Deviation Sum	Observed Frequency	Observed Probability	Expected Probability
0	16	.18	.04
2	49	.57	.19
4	18	.21	.29
6	3	.03	.29
8	1	.01	.19

$\chi^2 = 145.07$, df = 4, $p < .001$

Table 5

Intercorrelations (r) of Selection Formula Factors plus Empathy Scale

	Interview	Empathy Scale	GPA	Otis
Interview	---			
Empathy Scale	.25*	---		
GPA	.23*	-.04	---	
Otis	.26*	.22*	.46**	---
Nelson-Denny	.12	.04	.38**	.64**

df = 85 (df = 80 for correlations involving empathy scale)
* $p < .05$ ** $p < .001$

PRECIS

EVALUATION OF TEACHING

MODERATOR: Arthur I. Rothman, Ed.D.
University of Toronto

HOW MEDICAL STUDENTS USE OBJECTIVES

This study examined (1) how medical students use objectives as they work through the curriculum, (2) what factors influence their use of objectives, (3) whether differences in use of objectives are related to students academic ability. The results, based on a 77% questionnaire return from one class of 58 students in an objectives-based medical curriculum, are reported and discussed.

WHAT SHOULD STUDENTS EVALUATE? A COMPARISON OF FACULTY AND STUDENT PERCEPTIONS

Faculty and students rated the appropriateness of sixteen topics in a student evaluation form. Despite differences between student and faculty perceptions and differences between basic science faculty and clinical faculty, it was possible to identify the most salient topics to be included.

TEACHING EFFECTIVENESS: A COMPARISON OF STUDENT RATINGS AND GRADUATE EFFECTIVENESS

Four years of accumulated physician assistant student ratings of course relevance and course quality are compared with an evaluation of graduate performance in the practice setting. The initial findings were that students tend to rate the relevance of certain courses significantly higher and the quality of certain courses significantly lower than graduate performance in the practice setting would indicate.

COMPARISON OF FACULTY AND CHAIRMEN PERCEPTIONS ABOUT PROMOTION CRITERIA

This is a study of the similarities and differences between the ratings by medical school faculty and department chairmen about which activities contribute most toward promotion. Faculty saw research and publication as the most important activities; chairmen had a broader perception and included research, teaching, patient care and administration-service in their list of high status activities.

HOW MEDICAL STUDENTS USE OBJECTIVES*

Terrill A. Mast, Ph.D.; George P. Evans, M.S.;
Reed G. Williams, Ph.D.; David L. Silber, M.D.
Southern Illinois University School of Medicine

The persuasive nature of Mager's[1] treatise on instructional objectives stimulated an array of educational developments and experimentation. The primary focus of this research[2-7] has been on the effects of objectives on student learning, retention and attitudes. Its primary vehicles have been instructional units varying in length from several minutes to a semester or more. Most studies have described how objectives were developed and utilized by faculty and have assumed that students in the experimental groups have used the objectives in the way faculty intended. While acknowledging that the advantages of objectives may be due as much to their effects on faculty as to their direct effects on students, the present research examines when and how medical students actually use objectives.

The Southern Illinois University School of Medicine undergraduate curriculum is entirely objectives-based.[8] The SIU program is divided into Sequence I, 12 months of basic sciences; Sequence II, six months of preclinical sciences; Sequence III, 12 months of hospital-based clinical clerkships; and Sequence IV, five months of primarily clinical electives. Objectives have been written for all learning modules of all four sequences, are given to students, and are available for public review.[9] These objectives are stated in terms of what students are expected to be able to do and represent all taxonomic levels. Frequent evaluations of students' mastery of the objectives determine their progress through the curriculum. All objectives, instructional activities, and measurement techniques are under periodic assessment and revision.

In the course of reviewing five years of experience in this objectives-based curriculum it became apparent that, while there was substantial positive feedback from students about the availability of objectives, there was no reliable information on how students who are experienced objectives-users actually use them, either in the SIU experience or from the reports of others. In this project complementary interview and questionnaire methods were employed to study (1) how the Class of 1978 reported using objectives as they worked through the curriculum, (2) what factors influenced their use of objectives, and (3) whether differences in reported use of objectives were related to differences in their academic abilities.

Method

A questionnaire employing open-ended and scaled items was mailed to the entire class of 58 students their final month of medical school. The questionnaire content was based upon the results of small group interviews about the research questions with twelve students from the class. Pre-medical undergraduate grade point averages and Medical College Admission Test (MCAT) scores were used as indices of medical school preparation. Part I and II NBME Subject Test scores were used as measures of student achievement. All three were correlated with questionnaire responses.

---

*Requests for reprints should be addressed to Terrill A. Mast, Office of Curriculum Affairs and Educational Resources, Southern Illinois University School of Medicine, P.O. Box 3926, Springfield, Illinois 62708.

Results

Questionnaire mailings yielded a 77% response rate (n=47).

To examine how much students used objectives in each stage of the curriculum, they were asked to rate, "How USEFUL were goals* to you in the average module in each sequence?" on a scale of 1-5 where 1 was "Detrimental," 2 was "No Effect," and 5 was "Essential." The modal responses were 5 in Sequence I, 4 in Sequence II, and 3 in Sequence III (the clerkship year). All students reported that Sequence I goals were of some use (ratings of 3,4, or 5) and 25 of 45 rated them as "Essential." Sequence II goals were rated as "Essential" by only six students out of 45. There were considerable differences among Sequence III clerkships, ranging from one clerkship in which 70% of the students rated objectives "Essential" to another in which 64% rated the objectives as having "No Effect." On another item, individuals with lower MCAT Science scores indicated a preference that goals be given greater emphasis in Sequence III ($r=-.35$, $p<.05$) and Sequence IV ($r=-.34$, $p<.05$). (These correlations were obtained after omitting one student with an MCAT Science score 100 points below all others. The correlations were of larger magnitude before removing this score.)

In another question bearing on the issue of differing usefulness of objectives in different sequences, students were asked to list the curriculum units in which the goals were most useful and those in which the goals were least useful. Of the curriculum units cited as having the most useful goals 25 were in Sequence I, 13 in Sequence II, and 17 in Sequence III. Of the curriculum units with the least useful goals 4 were in Sequence I, 4 were in Sequence II, 32 were in Sequence III and 8 were in Sequence IV.

Three open-ended questions examined the role of objectives in students' study patterns. In Sequence I, 89% of the respondents indicated a step-by-step study pattern in which they read the objectives prior to reading the assigned resource material. Some students also volunteered the information that they wrote out answers to the objectives or belonged to a study group whose members wrote and exchanged answers to objectives. In Sequence II, 74% reported study patterns in which they read the goals first. Equivalent data were not obtained for Sequence III and IV, but the interviews had suggested that students use goals less in Sequence III and IV than in either of the first two sequences.

Two questions with multiple parts examined the factors which influenced student use of objectives. The first asked how various factors mentioned during the interviews tended to increase or decrease student use of objectives. The results (presented in Table 1) showed that the factor which tended to increase student use of objectives most was the expectation that they would be tested on the goals. A related item showed that use of goals also varied directly with the frequency of testing. A second cluster of variables which tended to increase use of objectives was the ease with which students could use the objectives. This is exemplified by the finding that clarity, conciseness, and a good match between the objectives and the reading material tended to increase student use of objectives. Conversely, having only a few broad, general objectives tended to decrease the use of objectives.

---

*The word "goals" is consistently used by SIU students to refer to instructional objectives and was used in the questionnaire. Throughout this article the two terms will be used synonomously.

The second multi-part question asked students to rate their agreement with several statements about objectives. The statements and the results are presented in Table 2. Students felt that some content areas were better suited to goals than others. Most students cited goals as promoting efficient use of study time. Goals were perceived as useful for promoting factual learning ($\bar{x} = 4.11$) but as having little effect on conceptual learning ($\bar{x} = 3.26$) or integration of knowledge ($\bar{x} = 3.04$). Twenty-five of forty-seven students agreed that goals primarily helped them pass tests. Students with lower MCAT Quantitative subtest scores tended to agree more strongly with this statement ($r=-.34$, $p<.05$).

The two items examining student perceptions of faculty compliance with the principle of testing only over objectives show that students thought there was fairly good compliance in Sequence I and moderate compliance in Sequence II. (The preliminary interview had shown that students perceived the compliance in Sequence III and IV as quite low.)

Finally, students were asked whether objectives inhibited their learning in any was and, if so, how. Of the 45 respondents to this question, 62% responded "no," 24% responded "yes," and 13% were ambivalent. The comments of students who reported that their learning was not inhibited by objectives suggested that if and when they found the goals to be inhibiting, they stopped using the goals. Comments of students reporting that their learning was inhibited by objectives suggested that they sometimes did not go beyond goals in their study, particularly when study time was very limited. These students reported that limiting their study to answering the goals caused them to bypass additional reading which was necessary for understanding the concepts and thereby resulted in a decreased appreciation of some subject areas, particularly among the basic sciences.

Of all the correlations between the questionnaire items and student entering GPA's, MCAT's and Parts I and II NBME scores, there were only seven statistically significant correlations, four of which dropped below significance when the one student with an extremely low MCAT Science score was omitted. Those which were significant have been reported above.

Discussion

The primary finding of this study is that students reported instructional objectives to be essential early in the curriculum, and to gradually decrease in usefulness throughout the four sequences. There are at least three possible explanations for this finding: (1) as student sophistication with the content of the medical curriculum develops, students have less need for objectives to help them identify the important information; (2) the quality of objectives is poorer in later sequences which made those objectives less useful; and (3) the objectives are most useful in acquiring cognitive knowledge and therefore less useful in later sequences where there is less emphasis on cognitive learning.

All the results of this questionnaire study are consistent with the view that objectives serve primarily to key students to the demands of the system. Objectives were used more when students expected them to correspond closely to examinations, when examinations were frequent, when students were under time pressure and there was greater need for efficiency, and when objectives were perceived as easily usable. There was little evidence that student perception and use of objectives was systematically related to student academic ability. The finding that objectives were less useful in Sequence II than in Sequence I reasonably follows from the students' observation that testing demands are somewhat less consistent with objectives in Sequence II than in Sequence I. The interview sessions and the "grapevine" indicate that the clerkship (in which objectives had the highest usefulness rating was the one whose tests were most directly consistent with objectives, and the clerkship with the lowest rating for its objectives was the one in which the objectives were of very little help in predicting test content.

While examining the usefulness of the objectives their quality is also brought into focus. Students do not use objectives perceived to be of poor quality. The medical students and Mager[1] both agree that objectives should be clear and concise to communicate the writer's intent. The students have added additional operational components to Magers' definition of a good objective by stating that useful objectives should also be narrowly focused, require only a short answer, and be sequentially and structurally matched to the resource material.

An unsettling finding in an entirely objectives-based curriculum is that students perceive objectives to be less useful in promoting the higher taxonomic levels of learning. This implies either that objectives do not increase efficiency of study for higher order learning or that the system does not demand (i.e., does not test) higher order learning. Although conscious efforts are made by SIU faculty and educational staff to assure that all levels of learning are examined, systematic review of SIU objectives and test items might well yield results similar to those reported by Gjerde,[10] a preponderance of both lower level objectives and test items. There may also, however, be an interaction between objective level and student standards for quality of objectives. If higher order objectives tend to be longer, less clear, require longer answers, and correspond more abstractly with the instructional material than lower order objectives, then students may perceive them as less useful regardless of their consistency with test demands.

The finding that some of the students felt inhibited by objectives bears examination. Those students indicated that when time was short and pressures great, they sometimes limited their study to those areas specified by the goals, and felt that this caused them to miss things they would have gained from doing additional reading. This clearly would be a potential danger in a system in which the goals inadequately specify the important learning whether those goals are explicit or implicit. If the goals do adequately specify what is important (i.e., relevant, required, necessary) and if the goals are achieved, then there can be no basis for the statement that goals inhibit learning. The real inhibitor is the universal competition for time in medical school which interferes with the achievement of goals.

These results imply 1) caution in the unquestioned advocacy of similar forms of objectives throughout the curriculum, 2) continued attention to the quality of objectives, and 3) continued attention to the adequacy of objectives at higher taxonomic levels. However, the questions raised by this study flow more directly than the curricular implications. Why are objectives less useful in later curriculum sequences? Do students use faculty objectives less in clinical years because they have internalized basic objective forms which they apply to all medical problems? Do students stop using faculty objectives as they develop independence, career direction, self-motivation and self-confidence? Are objectives less useful in promoting higher taxonomic levels of learning, or are higher taxonomic levels less tested? Why are there no correlations with the limited achievement and aptitude measures available? When time is adequate, and the quality and scope of objectives are adequate, are objectives ever perceived to inhibit learning? Can all important learning be measured, and can all important learning be specified by objectives? Further research will attempt to answer such questions.

REFERENCES

1. MAGER, R.F. Preparing Instructional Objectives. Palo Alto: Fearon, 1968.
2. DUCHASTEL, P.C. and MERRILL, P.F. The Effects of Bahavioral Objectives on Learning, Review of Educational Research, 43(1): 53-69, 1973.
3. TENNYSON, C.L. and TENNYSON, P.D. Instructional Objectives - Who Needs Them? Improving Human Performance Quarterly, 4(1): 12-16, 1975.
4. HARTLEY, J. and DAVIES, I.K. Preinstructional Strategies: The Role of Pretests, Behavioral Objectives, Overviews and Advance Organizers. Review of Educational Research, 46(2): 239-266, 1976.
5. MELTON, R.F. Resolution of Conflicting Claims Concerning the Effect of Behavioral Objectives on Student Learning. Review of Educational Research, 48(2): 291-302, 1978.
6. FAW, H.W. and WALLER, T.G. Mathemagenic Behaviors and Efficiency in Learning from Prose. Review of Educational Research, 46(4):691-720, 1976.
7. BLOCK, J.H. and BURNS, R.B. Mastery Learning. Review of Educational Research, 4: 3-49, 1976.
8. SILBER, D.L., WILLIAMS, R.G., PAIVA, P.A., TAYLOR, D.D., and ROBINSON R. The SIU Medical Curriculum: Systemwide Objectives-Based Instruction. Journal of Medical Education, 53(6): 473-479, 1978.
9. SOUTHERN ILLINOIS UNIVERSITY SCHOOL OF MEDICINE. Curricular Objectives 1976.
10. GJERDE, C. The Domains and Cognitive Processes in Medical School Objectives. Journal of Medical Education. 53(4): 352-355. 1976.

TABLE 1

RANK ORDER OF FACTORS INFLUENCING STUDENT USE OF OBJECTIVES*

Factors	Mode(s)	Mean	S.D.
Expectation that goals would be tested	5	4.57	0.81
Clear, concise goals	4,5	4.20	0.96
Goals match material well	5	4.18	1.19
Frequent testing	5	4.02	1.04
Goals requiring short answers	4	3.89	0.95
Shortage of time	5	3.74	1.24
Expectation of multiple choice tests	3	3.04	1.01
Prior knowledge of material	3	2.78	0.89
Large number of goals	3	2.67	1.16
Interest in the material	1,2,3	2.35	1.10
Few broad goals	2	1.93	1.12

* Mean ratings of 47 students on a scale of 1 = DECREASED to 5 = INCREASED.

TABLE II

AGREEMENT WITH STATEMENTS ABOUT OBJECTIVES*

Statements	Mode	Mean	S.D.
You were tested only over the goals in Sequence I.	4	3.96	1.06
You were tested only over the goals in Sequence II.	4	3.38	1.01
There should be greater emphasis on goals in Sequence III.	1	2.68	1.38
There should be greater emphasis on goals in Sequence IV.	1	2.00	1.25
Goals helped improve your learning of factual material.	4	4.11	0.89
Goals helped improve your conceptual understanding of material.	3	3.26	1.03
Goals promoted efficient use of your study time.	4	4.17	0.89
Goals primarily helped you pass tests.	4	3.53	1.08
Some subjects lend themselves to goals	5	4.64	0.56
You usually went beyond the goals.	4	3.63	0.82
The goals helped you integrate.	3	3.04	1.02
Having goals reduced your test anxiety.	4	3.66	1.23
If you teach, you will write goals.	4	3.85	0.95
Using goals has helped prepare you for independent learning.	4	3.38	1.21
You set your own goals in Sequence IV.	4	3.31	1.20

* Mean ratings of 47 students on a scale of 1 = DISAGREE to 5 = AGREE.

WHAT SHOULD STUDENTS EVALUATE?
A COMPARISON OF FACULTY AND STUDENT PERCEPTIONS

Michael B. Donnelly, Ph.D., Carolyn Puhl, Daniel S. Fleisher, M.D.
Jeffrey H. Garrison, M.D., and Stanley Kaplan, Ph.D.
The Medical College of Wisconsin

Ratings by students, while only partial indicators of teaching effectiveness and course quality, have been shown to be rather valid and reliable[1]. If they are to be utilized, it would appear to be critical to first determine what topics are important for students to rate in evaluating the instructor and the course.

In a thorough review of the literature, Kulik and McKeachie[1] reported four factors that represented what students rated most often in evaluating their instruction. These were: 1) Skill; 2) Rapport; 3) Structure; and 4) Overload. Closely related to these factors were the three dimensions that Slotnick and Curkovic[2] found in medical students' evaluation of their basic science instruction, namely 1) faculty-student rapport; 2) outside work or overload; and 3) aspects of student comprehension. In a study of third and fourth year medical students Stritter, Hain and Grimes[3] discerned six factors. Their factors #1 and #2 appear to be closely related to the aforementioned factor of faculty-student rapport and their factor #4 seems to be the same as Kulik and McKeachie's factor concerning structure. Their three other factors are limited to medical students in a clinical setting.

This study, utilizing the factors of student evaluation noted in the literature, was designed to discover the characteristics and differences, if any, in what students and faculty perceive to be relevant topics for evaluation by medical students. They were carried out as part of an effort by a faculty committee to revise an existing student evaluation form. The resultant data were used by the committee in creating a revised form.

Method

A questionnaire concerning student evaluations was sent to all faculty and students of our medical college. Although the questionnaire covered policy issues (i.e. who should receive the results of student evaluations, in what format should the results be presented, etc.), in addition to "what should be evaluated" issues, only the latter will be the subject of this paper.

Twenty topics derived from the aforementioned factors in the literature were rated each on a seven point scale, from Not Appropriate to Most Appropriate. One hundred forty-one faculty and one hundred fifty-six students returned the completed questionnaires. No claims are made about the completeness or random nature of the sample. However, we have assumed the returns represent data from those who have rather clear opinions on the utility of student evaluations.

The mean responses to each of the twenty items were compared for the following

Reprint Requests: Michael B. Donnelly, Ph.D., Office of Curricular Affairs,
The Medical College of Wisconsin, 8701 Watertown Plank Road,
Milwaukee, Wisconsin 53226

groups: 1) Faculty vs. students; 2) students in first two (basic science) years vs. students in the last two (clinical) years; 3) basic science faculty vs. clinical faculty; and 4) junior faculty vs. senior faculty.

Knowledge of the mean ratings and the standard deviations alone did not provide sufficient understanding of the homogeneity underlying these mean ratings. For this purpose, an agreement index was derived for each group's responses to all twenty items using the following definition:

$$\text{Agreement Index} = \left(1 - \frac{SX^2}{SR^2}\right) \times 100$$

where SX^2 = Observed Variance

SR^2 = Random Variance

The limits of this index were defined as follows: 1) if the ratings are randomly distributed across the continuum, the agreement index will equal 0%; 2) if there is complete agreement (SX = 0.00), then the index will equal 100%.

Results

1) Faculty vs. Students

On all but 2 items (i.e. 2A and 4D) in Table 1, student responses were significantly different and higher than faculty responses. However, the correlation of mean faculty and student responses was 0.90 indicating that these two groups rank ordered the importance of these items in very similar ways.

Table 2 shows that there was more agreement among student ratings than there was among faculty ratings (17 vs. 2 with one tie) (Sign Test - $p < .01$).

If one arbitrarily defines low agreement as any index less than 50%, then faculty tended not to agree on 13 of the 20 items (e.g. 2B, 3A, 3B, 4B, 4E, 4F, 4H, 4I, 4J, 4K, 4L, 4O and 4P). On the other hand students tended not to agree on 5 of the 20 items (e.g. 3A, 3B, 4F, 4O and 4P).

Table 3 shows the grouping of faculty and student responses (i.e. to items 4A through 4P) into priority levels. Although some differences exist between the contents of each level, there are more similarities than differences. Given the fact that 4.0 was the midpoint of possible responses, however, faculty felt that four items (4B, 4E, 4I and 4O) were inappropriate for students to evaluate while students did not feel that it was inappropriate for them to evaluate any of the measures listed. It should be noted that the means presented in Table 3 differ slightly from the means in Table 1, the differences being a function of how the two computer programs handle missing data.

2) Students in Basic Science Years vs. Students in Clinical Years

Table 1 indicates significant differences only in items 4F and 4G, not surprisingly rated higher by students in their clinical years.

3) Basic Science Faculty vs. Clinical Faculty

Table 1 shows that basic science faculty rated 3 items (4J, 4M and 4N) significantly higher than clincal faculty while the reverse was true for 7 items (3A, 4B, 4E, 4F, 4G, 4H and 4O).

181

4) Junior Faculty vs. Senior Faculty

Table 1 shows that these two groups differed significantly only on two items (3A and 4E) with senior faculty rating these items more positively.

In order to summarize the results, a multiple discriminant function analysis was performed comparing basic science faculty, clinical faculty, students in basic science years and students in clinical years. Two significant discriminant functions ($X^2 = 211.9$, $p < .01$ and $X^2 = 63.9$, $p < .01$) were obtained. The first function clearly separated student and faculty ratings in a bipolar fashion. The second function differentiated basic science faculty and clinical faculty with students at the center of this function. No significant differentiation of students was obtained. Thus, the discriminant function indicated that difference between faculty and students was the most salient dimension of comparison.

Discussion

If the major goal of student evaluations is to provide useful information to faculty about student perceived strengths and weaknesses of a course, it seems logical to assume that this possibility is maximized when faculty and students agree on what is to be evaluated. This study has demonstrated that there are large and consistent differences between faculty and students. Students tend to see all the traits as more important for evaluation. In fact, they believe that all the items listed are appropriate for their judgement. On the other hand, faculty clearly differentiate variables which they feel should be evaluated from those which should not. The problem then is how to solve this dilemma. Since students and faculty rank order items in very similar ways, it would appear fair to limit the student evaluations to those topics in which faculty and students agree in their relative rankings. This study has indicated, though not in an exhaustive manner, what those variables might be.

The comparison of basic science and clinical science faculty indicated that there are differences in what they would like students to evaluate. It should be noted, however, that similar differences were not reflected in the ratings of students in the basic science years and students in their clinical years. At any rate, it appears that basic science faculty are relatively more concerned with students evaluating faculty availability and student time and satisfaction; while clinical faculty are more concerned with students evaluating both student and faculty competence. It is probable that these ratings reflect the different roles of basic science and clinical science faculty in schools of medicine. However, these data also suggest that we might well design different student evaluation forms for evaluating basic science and clinical science faculty.

The agreement ratings reflect the fact that the students tend to agree on what is important to rate. This is probably a function of the fact that, in a rather undifferentiated view, students perceive that all elements in the educational process are important to evaluate. Individual faculty members have developed a rather complex view of what to evaluate. The ideographic nature of these perceptions probably leads to the relatively low level of agreement between faculty.

Finally, it appears clear that student evaluations forms should stress the process of the educational encounter but should not be concerned with defining what should be taught or what defines competency of the teachers.

Conclusions

Faculty and students believe that students should evaluate a fairly large number of items relating to the instructional process with students being more positive and less discriminating than faculty. In addition, student ratings show significantly more agreement than do faculty ratings.

Some minor differences in responses to what students should evaluate exist between junior and senior faculty and between students in their first two years and those in their last two years of medical school.

Moderate differences in responses were noted between basic science faculty and clinical faculty.

It would appear that knowledge of the degree and nature of faculty-student differences in what students should evaluate is critical in the design of student evaluations that are likely to lead to improvements in the instructional process.

References

1. Kulik, James A., and McKeachie, Wilbert J.: The evaluation of teachers in higher education. In F.N. Kerlinger Review of Research in Education, Vol. 3, Itaska, Ill. F.E. Peacock, 1975.

2. Slotnick, Henry B., Durkovic, Russell G.: Dimensions of medical students' perceptions of instruction. J. of Med. Ed., Vol. 50, 662-665, July 1975.

3. Stritter, Frank T., Hain, Jack D., and Grimes, David A.: Clinical teaching reexamined. J. of Med. Ed., Vol. 50, 876-881, Sept. 1975.

Table 1 - DIFFERENCES IN RATINGS BETWEEN VARIOUS GROUPS

	Mx Faculty	Mx Students	F	Mx Basic Science Students	Mx Clinical Students	F	Mx Basic Science Faculty	Mx Clinical Faculty	F	Mx Junior Faculty	Mx Senior Faculty	F
Q2A. Course development and change	4.91	4.92	0.00	4.93	4.87	0.14	4.82	5.00	0.42	4.78	5.01	0.89
Q2B. Recommendations for promotion and tenure of faculty	3.50	4.43	31.99*	4.45	4.42	0.03	3.30	3.59	0.68	3.98	3.54	0.05
Q3A. Evaluating what they should have learned	3.50	3.87	4.82*	3.93	3.76	0.56	2.94	3.65	6.01*	3.74	3.27	3.96*
Q3B. Evaluating the effectiveness of individual instructors	4.20	4.72	10.68*	4.67	4.77	0.19	3.94	4.27	1.49	4.25	4.18	0.09
Q4A. Individual faculty teaching skills	5.21	6.07	47.61*	6.00	6.14	0.80	5.24	5.20	0.03	5.25	5.20	0.04
Q4B. Course content	3.62	4.78	50.96*	4.74	4.82	0.15	3.00	3.73	6.16*	3.79	3.47	1.61
Q4C. Course evaluation	4.65	5.65	54.87*	5.58	5.69	0.38	4.48	4.67	0.54	4.70	4.61	0.18
Q4D. Clarity of course objectives	5.37	5.60	2.08	5.47	5.71	1.23	5.24	5.43	0.47	5.41	5.34	0.10
Q4E. Clinical relevance of basic science	3.43	4.21	21.27*	4.13	4.24	0.23	2.70	3.66	9.69*	3.79	3.23	4.59*
Q4F. Clinical relevance of clinical science	3.97	4.59	13.60*	4.29	4.83	6.44*	3.36	4.15	6.56*	4.13	3.87	1.04
Q4G. Development of clinical problem solving skills	4.57	5.03	9.87*	4.74	5.24	6.77*	4.06	4.70	7.08*	4.72	4.46	1.53
Q4H. Importance of content to professional goals	3.89	4.84	36.02*	4.71	4.92	0.97	3.39	3.98	4.17*	3.95	3.89	0.06
Q4I. Amount of time allotted to course	3.68	5.15	78.36*	5.08	5.25	0.65	3.73	3.62	0.11	3.90	3.58	1.56
Q4J. Overall student satisfaction	5.47	6.32	32.07*	6.34	6.26	0.20	5.94	5.36	4.11*	5.59	5.39	0.71
Q4K. Meeting of psychological needs of students	4.99	5.83	23.22*	5.79	5.81	0.00	5.09	5.01	0.06	5.23	4.88	1.73
Q4L. Development of scientific skills	4.22	4.88	18.93*	4.93	4.79	0.57	3.97	4.21	0.75	4.46	4.11	2.25
Q4M. Faculty availability	5.73	6.17	10.91*	6.20	6.18	0.01	6.12	5.56	4.90*	5.51	5.80	1.77
Q4N. Use of student time	4.91	5.85	35.34*	5.83	5.86	0.02	5.45	4.69	7.28*	4.82	4.90	0.12
Q4O. Faculty professional competence	3.50	4.84	57.51*	4.96	4.69	1.36	2.64	3.69	11.85*	3.79	3.35	2.69
Q4P. Sequencing of course content	4.21	4.81	12.16*	4.84	4.79	0.06	3.85	4.34	2.52	4.18	4.28	0.15

* p < .05

TABLE 2

Faculty and Student Agreement in Rating Topics

TOPICS	FACULTY Measure of Agreement	STUDENTS Measure of Agreement
Q2A. Course development and change	53%	69%
Q2B. Recommendations for promotion and tenure of faculty	24%	60%
Q3A. Evaluating what they should have learned	38%	38%
Q3B. Evaluating the effectiveness of individual instructors	48%	37%
Q4A. Individual faculty teaching skills	67%	76%
Q4B. Course content	48%	53%
Q4C. Course evaluation	63%	67%
Q4D. Clarity of course objectives	56%	54%
Q4E. Clinical relevance of basic science	38%	50%
Q4F. Clinical relevance of clinical science	38%	48%
Q4G. Development of clinical problem solving skills	56%	59%
Q4H. Importance of content to professional goals	46%	56%
Q4I. Amount of time alloted to course	42%	54%
Q4J. Overall student satisfaction	47%	70%
Q4K. Meeting of psychological needs of students	34%	54%
Q4L. Development of scientific skills	48%	61%
Q4M. Faculty availability	63%	73%
Q4N. Use of student time	51%	56%
Q4O. Faculty professional competence	38%	44%
Q4P. Sequencing of course content	41%	45%

TABLE 3

Significant Differences Between Items To Be Evaluated By Students

FACULTY

ANOVA

Source	df	MS	F
ITEMS	15	82.50	39.73*
ERROR	2130	2.08	

STUDENTS

ANOVA

Source	df	MS	F
ITEMS	15	63.88	37.58*
ERROR	2376	1.70	

DIFFERENCE LEVELS**

	FACULTY	STUDENTS
Level 1	Q4M. Faculty availability (5.77)*** Q4J. Overall student satisfaction (5.50) Q4D. Clarity of course objectives (5.40)	Q4J. Overall student satisfaction (6.37) Q4M. Faculty availability (6.21) Q4A. Individual faculty teaching skills (6.11) Q4K. Meeting psych. needs of students (6.00)
Level 2	Q4A. Individual faculty teaching skills (5.24) Q4K. Meeting psych. needs of students (5.07) Q4N. Use of student time (4.96)	Q4N. Use of student time (5.90) Q4C. Course evaluation (5.71) Q4D. Clarity of course objectives (5.66)
Level 3	Q4C. Course evaluation (4.68) Q4G. Development of clinical P.S. skills (4.66)	Q4I. Amount of time allotted to course (5.21) Q4G. Development of clinical P.S. skills (5.15) Q4L. Development of scientific skills (4.91) Q4H. Importance of content to professional goals (4.89) Q4O. Faculty professional competence (4.87) Q4P. Sequencing of course content (4.83) Q4B. Course content (4.79)
Level 4	Q4L. Development of scientific skills (4.24) Q4P. Sequencing of course content (4.23) Q4F. Clinical relevance of clinical science (3.97) Q4H. Importance of content to professional goals (3.88)	Q4F. Clinical relevance of clinical science (4.65)
Level 5	A4I. Amount of time allotted to course (3.68) A4B. Course content (3.61) Q4O. Faculty professional competence (3.49) Q4E. Clinical relevance of basic science (3.40)	Q4E. Clinical relevance of basic science (4.22)

* $p < .01$

** Defined by Newman-Keuls Post Hoc Test.

*** Numbers in parentheses are mean ratings.

TEACHING EFFECTIVENESS: A COMPARISON OF STUDENT RATINGS AND GRADUATE EFFECTIVENESS*

Roger A. Lanier, Ph.D., Byron Williams, P.A.-C., and Bruce R. Niebuhr, Ph.D.
The School of Allied Health Sciences
The University of Texas Medical Branch at Galveston

Introduction

The debate as to what constitutes teaching effectiveness has continued for more than four decades. The debate generally surrounds whether the evaluation of several "process" variables are a sufficient measure of teaching effectiveness or whether the evaluation of "product" variables are necessary.

Recently, institutions of higher education have been increasingly criticized for their lack of attention to teaching effectiveness.[1] As a result, some institutions are attempting to construct promotion, tenure, and salary systems which are directly related to teaching effectiveness. Therefore, while the debate has raged for decades, there is currently an urgency for higher education, and medical or health education, specifically, to clarify the teaching effectiveness issues and facilitate a critical examination of their educational missions.

The present study compares the results of several years of student ratings of courses with the graduate's effectiveness in the health setting following graduation. The implications and significance of the present discrepancies for medical education are noted and discussed.

Review of the Literature

About eighteen years ago, student ratings of teaching were cited as the most frequent method of evaluating teaching effectiveness.[2] However, a substantial decline in the use of student ratings resulted about twelve years ago.[3] One explanation for the decline may be that of all the methods employed to evaluate teaching effectiveness, perhaps the most controversial is the use of student ratings.[1]

Possibly, much of the controversy is the result of the dubious nature of many of the educational process oriented variables and their attendant measurement problems. For instance, the following instructions have been provided for the rating of an "ideal teacher".[4]

> "...A person deserving the highest possible rating, as a teacher, was described as one who is both capable and efficient, who loves his job and manages to inspire his students, who is himself inspired with his work, who is talented, and who not only respects and appreciates his students, but also has good relations with them. The highest rating as a personality was to be given to a very healthy, well integrated person, subjectively at ease with himself, happy or content, using all his constructive capacities, enjoying

---

*Requests for reprints should be directed to Roger A. Lanier, Ph.D., Assistant Dean of Academic Affairs, The School of Allied Health Sciences, The University of Texas Medical Branch, Galveston, Texas 77550.

> life without neurotic or psychotic maladjustment."

In a recent literature review,[5] approximately twenty-one studies were cited as suggesting more than twenty-one factors which were recommended for rating teaching effectiveness. Some of the factors listed were the extent to which teachers were skilled at instruction; well prepared for class; provided clear explanations; stressed important material; used appropriate examples; inspired confidence in their knowledge of the material; motivated students; provided an interesting or enthusiastic presentation; stimulated intellectual curiosity or critical thinking; and provided new applications for the material. Other factors reported were establishing rapport with students; being sympathetic to students' reactions; being friendly; being flexible; providing feedback to students; presenting a pleasing personality; having a sense of humor; being sincere; having broad background interests; and being fair.

The reliability and validity of such student ratings of teaching effectiveness have also been examined. The reliability of student ratings appears to be adequate when measured at either two points in the semester or within the same administration of the instrument. Several studies have been cited[5] as suggesting that student ratings are both internally consistent and stable.

The primary question, underlying much of the debate, appears to be the validity of the measures for the assessment of teaching effectiveness. Even the early rationales for utilizing student ratings seemed to suggest that they were readily available proxy measures. One rationale suggests that since there is so little agreement on what constitutes teaching effectiveness, the best subjective judgment can be provided by the student "consumers" of the course.[6,7] Another rationale suggests that the measurements are readily available; since students spend more time with the teacher than anyone else, observing teaching from the client's point of view.[1]

Validity studies, within this framework, have generally remained within the classroom; where the results have been mixed. For instance, twenty-nine studies have been cited[5] as examining the relationship between student ratings and student grades. Fifteen of these studies found no relationship; thirteen found a positive relationship; and one found a negative relationship. Two studies have been cited[5] which found a positive relationship between students who thought they had achieved course objectives and students who received high grades. Additionally, two studies were cited[5] which found a positive relationship between students who rated teachers high on various personality factors and students who received high grades.

An alternative rationale has been that the appraisal of what students learn is the most direct and "ultimate" evaluation of teaching effectiveness.[8,9] Within this framework, there are at least two significant limitations associated with student ratings of teachers,[1] which are: (1) Beginning or uniformed students are in no position to judge the accuracy or validity of course content, and (2) Student rejection of certain content as irrelevant or useless may be erroneous. When this rationale has been related to current student ratings, the best summary may be that:[10]

> "...the present data indicate that students are less than perfect judges of teaching effectiveness if the latter is measured by what they have learned. If how much students have learned is considered

> to be a major component of good teaching, it must be concluded that good teaching is not validly measured by student evaluations in their current form."

Unfortunately, only a limited number of empirical investigations have examined student ratings compared with the student's feelings or performance as graduates. One study has compared what alumni ten years out of college feel are important qualities of teaching (adequacy of preparation, interest, stimulation of intellectual curiosity, and progressive attitude) with current undergraduate opinions and found a substantial agreement.[11] They also found a positive correlation between student ratings and alumni ratings of the same instructors. Another study found that students in an aircraft mechanics course at an Air Force installation made gains in information and in practical job sample performance which were significantly correlated with their overall ratings of the course (r = .32 for information and .39 for practical performance).[12]

The present study examines two dimensions of student ratings, relevance of the course and quality of the course; relative to how often the graduate performs the activity and how effectively they perform the activity in the health setting. Each dimension is compared to determine if there is a significant difference between the feedback students provide prior to entering the health setting and the activities they actually perform following graduation.

Methodology

Student course evaluations were obtained for the classes of 1975, 1976, 1977 and 1978 from the Physician's Assistant Program of The School of Allied Health Sciences, The University of Texas Medical Branch. Total enrollment for these four years was 72. In 1978 a follow-up study was conducted of the 46 individuals who graduated through 1976. These graduates had been in practice for over a year, a time period utilized to establish a stable baseline of activities which the physician's assistant would be performing. The study included gathering data on the percentage of time they spend in specific activities. To verify this data, the same information was collected from each employing physician, but this has been reported elsewhere.[13,14] Student ratings of teaching effectiveness were collected in every course and stored over a three year period. Within the evaluation instrument, the students were asked to: (1) rate the overall quality of the course, and (2) indicate how relevant was the course.

Relevance of an activity for the graduates was defined as follows: an activity utilized for more than 20% of patients seen was considered relevant. These graduate responses were compared to student ratings of course relevance through the use of a chi-square. The graduate's self-ratings of performance (good or poor quality) encompassed under each course content area were compared to student ratings of quality through the use of a chi-square.

Results and Discussion

Student and graduate ratings of course relevance were found to differ significantly for patient evaluation, laboratory procedures, pharmacology, and community health (See Table 1). Student and graduate ratings of course quality differed significantly for medical interviewing, anatomy and physiology, laboratory procedures, pharmacology, patient education, and community health (See Table 2). Thus, when student ratings of certain aspects of teaching effectiveness are viewed relative to certain graduate performance measures, the feedback received tends to be erratic.

The literature suggests that there is only a limited consensus as to what constitutes teaching effectiveness and there are primarily only subjective measures which reflect its existence. Students, however, provide convenient measures of something they observe more frequently than others and rate fairly reliably without the benefit of a background in the content, a formal understanding of teaching techniques, training in personality assessment, or substantial experience in the professional area.

The present data suggest that student ratings of the relevance may be independent of how often they perform certain activities in the practice setting. Students tended to think courses were more relevant than they actually were. In two courses, ratings actually reversed directions, e.g., relevant to not relevant.

The present data also suggest that student ratings of the quality of instruction may be independent of how effectively they perform certain activities in the practice setting. Many courses which received low ratings on the quality of teaching effectiveness produced graduates which were competent practitioners in the content area. This seems to suggest that while the quality of teaching effectiveness failed to meet student expectations, when the demands of the practice setting are considered, graduates were more competent than their rating of the course would indicate.

To what extent graduates became competent practitioners in spite of the course through "on the job" training, independent reading, or some other source was beyond the control of this study. These results should be viewed as tenative until future research can control for these additional influences. The significance of these tenative results, however, lies in viewing teaching effectiveness from the broader perspective of the professional practice setting.

Student ratings of teaching effectiveness may provide valuable feedback relative to certain aspects of the "process" of teaching and learning, if the proper questions are raised. However, student ratings appear to provide erratic feedback relative to certain aspects of the "product" of teaching and learning.

It appears that student ratings of teaching effectiveness may be necessary, but fail to be sufficient measures of teaching effectiveness. It seems that a measure of the educational product, or graduate performance, may not be a sufficient measure of teaching effectiveness, but it certainly seems a necessary measure.

The implications of reducing something as complex as teaching effectiveness to a convenient and partial, if not superficial, measure seems potentially disastrous. As these critical medical education decisions are confronted, the present study would recommend that they receive the concentrated attention and rigorous research which the decisions require.

REFERENCES

1. Rippey, R. M. Student evaluations of professors: Are they of value? *Journal of Medical Education*, *50*:951-956. 1975.

2. Gustad, J. W. Policies and practices in faculty evaluation. *Educational Record*, *42*:194-211. 1961.

3. Gustad, J. W. Evaluation of teaching performances: Issues and possibilities. In C. B. T. Lee (Ed.), *Improving college teaching*. Washington, D. C.: American Council on Education, 1967.

4. Maslow, A. H. & Zimmerman, W. College teaching ability, scholarly activity and personality. *Journal of Educational Psychology*, *17*:185-189. 1956.

5. Costin, F., Greenough, W. T., and Menges, R. J. Student ratings of college teaching: Reliability, validity, and usefulness. *Review of Educational Research*, *41(5)*:511-535.

6. McKeachie, W. J. *Teaching tips. A guidebook for the beginning college teacher*. (6th ed.) Lexington, Mass: D. C. Health, 1969.

7. McKeachie, W. J. Student ratings of faculty. *AAUP Bulletin*, *55*:439-444. 1969.

8. Tyler, R. W. The evaluation of teaching. In R. M. Cooper (Ed.), *The two sides of the log*. Minneapolis: University of Minnesota Press, 1958.

9. Cohen, A. M., & Brawer, F. B. Measuring faculty performance. Washington, D.C.: ERIC Clearinghouse for Junior College Information/American Association of Junior Colleges, 1969.

10. Rodin, M. and Rodin, B. Student Evaluation of Teachers. *Science*, *177*:556-569, 1973.

11. Drucker, A. J. and Remmers, H. H. Do alumni and students differ in their attitudes toward instructors? *Journal of Educational Psychology*, *42*: 129-143, 1951.

12. Morsh, J. E., Burgess, G. G., and Smith, P. N. Student achievement as a measure of instructor effectiveness. *Journal of Educational Psychology*, *47*:79-88, 1956.

13. Lanier, R. A., Scott, E., Williams, B., and Bing, R. K. Evaluation of Program Effectiveness Through Graduate and Employer Follow-up. Paper presented at The Evaluation Network Convention. Aspen, Colorado. August 14-16, 1978.

14. Lanier, R. A., Niebuhr, B. R. & Williams, B. Program Evaluation of a Physician's Assistant Educational Program. Paper presented at Seventh Annual Conference on Physician's Assistants, Hollywood, Florida, April 1979.

Table 1
Relevance of Program Content Areas: Student Evaluation and Graduate Utilization

Course Content Area	Respondents	N	% Responding		χ^2
			Relevant	Not Relevant	
Medical Interviewing	Students	52	90	10	.4
	Graduates	35	86	14	
Patient Evaluation	Students	50	100	0	6.2*
	Graduates	34	88	12	
Anatomy & Physiology	Students	56	89	11	1.0
	Graduates	27	82	18	
Laboratory Procedures	Students	69	89	11	35.4***
	Graduates	35	31	69	
Pharmacology	Students	70	92	8	4.0*
	Graduates	30	77	23	
Patient Education	Students	67	72	28	.2
	Graduates	37	68	32	
Community Health	Students	51	63	37	16.2***
	Graduates	30	23	77	

* $p < .05$ *** $p < .001$

Table 2
Quality of Instruction: Student Evaluation and Graduate Effectiveness

Course Content Area	Respondents	N	Quality (% Responding)		χ^2
			Good	Poor	
Medical Interviewing	Students	53	61	39	18.42***
	Graduates	39	100	0	
Patient Evaluation	Students	50	98	2	.79
	Graduates	39	100	0	
Anatomy & Physiology	Students	58	67	33	10.78**
	Graduates	39	33	67	
Laboratory Procedures	Students	71	71	29	9.82**
	Graduates	26	100	0	
Pharmacology	Students	69	41	59	12.24***
	Graduates	39	77	23	
Patient Education	Students	68	33	67	46.28***
	Graduates	39	100	0	
Community Health	Students	52	48	52	23.86***
	Graduates	31	100	0	

** $p < .01$ *** $p < .001$

COMPARISON OF FACULTY AND CHAIRMEN PERCEPTIONS
ABOUT PROMOTION CRITERIA

Craig Gjerde, Ph.D., Sandra Colombo, M.A.
The University of Michigan

Introduction

The faculty in a medical school are recruited by the departmental chairmen. Chairmen delegate service, teaching, and patient care responsibilities to their faculty. If a faculty member does well in his four areas of faculty responsibility (research, teaching, patient care, and administration-service), his chairman may choose to nominate him to the promotion committee as a candidate for promotion to a high academic rank. Such a system can work smoothly if the chairman and his faculty have common perceptions about the criteria for promotion. When the chairman and the faculty have different perceptions about what is important, there will be dissatisfaction. This is a study of the similarities and differences in the perceptions of departmental chairmen and medical school faculty about what activities are the most important criteria for promotion.

Methods

Over 100 interviews were conducted with faculty and department chairmen to determine their perceptions about the activities which contributed most toward promotion to a higher rank. On the basis of these interviews a questionnaire of sixty-nine activities was constructed. Forty-one activities were subsumable under the four major areas of faculty responsibility. Twenty-eight additional activities dealing with personal qualities and behavior (hereafter called other attributes) were suggested during the interviews.

Respondents were asked "How do the following activities apply as criteria of promotion in your department?" The response scale included: Detrimental, Undesirable, No Effect, Desirable, Essential, and Not Sure. In the analysis, the responses Detrimental through Essential were scaled as 1 through 5.

The mean rating given each activity by 393 faculty (80% return) and the 22 (100% return) chairmen was calculated and reported in Table 1. Significant differences between the means of the faculty and chairmen are indicated by asterisks next to the faculty mean. Analyses of variance was used to compare the means. Since there were sixty-nine comparisons made, three or four could be expected to turn out significant at the =.05 level. For this reason only differences whose chance likelihood was less than 0.01 will be discussed.

Reprints from Craig Gjerde, Ph.D., Learning Resource Unit, University of Iowa Medical School, Iowa City, Iowa 52242

This study was supported by a grant from the National Fund for Medical Education.

Results

Faculty and chairmen basically agreed in their ratings of fifty-seven of the sixty-nine activities. There were, however, some important differences in the ratings of certain activities. Faculty accorded high status (ratings of 4.0 and higher) to eleven activities; chairmen judged nineteen activities to be of high status. In the eyes of the faculty, research and its publication were the most important activities; eight of the eleven high status activities were research-related, and none were related to teaching or patient care. Chairmen included research, administration-service, teaching, and patient care in their list of high status activities; they had a much broader perception of the activities which contributed toward promotion.

In their perception of the promotion system, faculty placed much greater emphasis ($p<.01$) on publishing a certain number of articles each year. They also placed greater emphasis ($p<.01$) on departmental dealings--negotiating promotions when hired, being a member of an inner group in the department, and not challenging what the department chairman asks you to do; chairmen assigned these last three activities small negative ratings, denying, in effect, that they made much difference either way.

Departmental chairmen rated eight activities significantly higher ($p<.01$) than the faculty. Counseling students and participating on committees outside the university were considered moderately important by chairmen but of essentially low importance by faculty. The other six activities which chairmen rated significantly higher than faculty were activities which chairmen perceived to have high impact on promotion: do an excellent job of teaching, be an innovative teacher, teach graduate students or house officers, be professionally respected by departmental members, have strong letters of reference from recognized experts in your field, and provide competent patient care.

Discussion

Chairmen and faculty jointly recognized the importance of research, publication, and achieving professional respect in the department. Chairmen, however, indicated that the performance of other non-research activities was also important for promotion.

The areas of discrepancy in ratings were indicative of potential faculty-chairman conflicts. If a faculty member believed that some activities (e.g., serving on departmental committees and teaching medical students) did not contribute toward his promotion, he would be less willing to accept those activities and might feel that his chairman was giving him busy work. A chairman who felt that those activities were desirable or essential experiences would want to give a departmental member a chance to experience them.

One possible explanation for these differences in perception is that there was a two-level operation involved in promotions. At the first level, the department, the faculty member had to be able to show participation and quality in teaching, competence in patient care (where applicable), service on departmental committees, and, of course, research. Chairmen claimed that all these activities entered into their perceptions as being important and that at the departmental level they used these criteria in deciding whether to nominate a faculty member. A chairman could make such judgments quite informally because (1) he was aware of those activities, having assigned them, (2) he could directly observe the performance of the faculty member or ask other departmental members for advice, and (3) most of the activities were departmentally administered.

At the second level, the institution, research/publication activities were preeminent; the quality and quantity of publications were essentially what counted for promotion. The promotion committee might implicitly assume that the chairman had already judged the non-research activities and the socialization/professional respect activities and found them to be adequate. If faculty were aware only of the formal review at the institutional level, their perceptions would be that research was the sine qua non for promotion.

Whatever the cause for the differences of opinion, it appears that departmental chairmen should take some initiative in discussing the four areas of faculty responsibility with their faculty and clarifying the expectations regarding the criteria for promotion at the departmental and institutional levels.

Table 1. RANKING OF ACTIVITIES AS CRITERIA FOR PROMOTION BY FACULTY AND CHAIRMEN

Activity	Faculty (N=393) Mean+	Chairmen (N=22) Mean
Be the primary author of journal articles	4.66	4.86
Publish articles in referred journals	4.65	4.77
Publish articles that make a contribution to the field	4.47	4.55
Publish certain number of articles each year	4.30***	3.64
Present papers at professional meetings	4.24	4.43
Achieve national recognition for research	4.23	4.14
Be professionally respected by your chairman	4.22	4.32
Accept responsibilities delegated by your chairman	4.13	4.36
Be professionally respected by department members	4.12***	4.64
Conduct lab-oriented research	4.07	3.91
Generate dollars to support research efforts	4.07	4.05
Conduct replicable research	3.94*	4.36
Present papers by invitation at professional meetings	3.89	4.05
Have at least 3 years at the current academic rank	3.87	3.67
Maintain a professional image	3.82*	4.18
Serve on departmental committees	3.78	4.00
Have high visibility within the academic community	3.75	3.82
Teach medical students	3.74*	4.27
Have an influential chairman	3.74	3.48
Conduct clinically oriented research	3.72	3.64
Write textbooks	3.71*	3.95
Teach graduate students or house officers	3.70**	4.18
Have strong letters of reference from recognized experts in your field	3.69***	4.23
Be personally liked by your chairman	3.68	3.50
Be a member of a national review committee: e.g., NIH	3.68	3.91
Do an excellent job of teaching	3.68***	4.23
Conduct research which is supportive of the department's direction	3.68	3.67
Be a member of an editorial board	3.66	3.77
Have high visibility within your department	3.65	3.62
Be personally liked by department members	3.65	3.82
Have a pleasing personality	3.61	3.59
Get outside employment offer from institutions of similar prestige	3.60	3.50

+ Where 5=Essential, 4=Desirable, 3=No Effect, 2=Undesirable, and 1=Detrimental. For the one-way ANOVA's between chairmen and faculty * means p<.05, ** means p<.01, *** means p<.001. Due to the large number of comparisons, the reader is advised to disregard the differences where a single asterisk (*) is shown.

Table 1 (continued)

Activity	Faculty (N=393) Mean	Chairmen (N=22) Mean
Profess congruence with departmental goals and norms	3.60	3.73
Receive teaching awards(s)	3.60	3.82
Serve on medical school committees	3.60	3.77
Be an innovative teacher	3.58***	4.05
Get favorable student evaluations for teaching	3.58*	3.91
Achieve national recognition for teaching	3.56*	3.86
Participate on university committees	3.55	3.77
Provide competent patient care	3.54**	4.05
Be a good administrator	3.53	3.62
Teach formal courses	3.52*	3.82
Have high visibility within the medical school	3.52	3.59
Follow regular channels to solve problems in your department	3.50*	3.82
Participate in grand rounds	3.43	3.68
Participate in interdisciplinary projects or programs	3.43	3.68
Be a course director	3.42	3.52
Achieve national recognition for patient care	3.42*	3.71
Conduct ward rounds	3.41	3.64
Don't challenge what the department chairman tells you to do	3.39**	3.29
Conform to the social norms of the department	3.39	3.29
Teach practicing physicians	3.37	3.62
Be a section chief	3.36	3.18
Be aggressive	3.35	3.14
Socialize with others in the department	3.33	3.29
Be advised by a senior faculty member on getting through the system	3.29*	2.90
Be a member of an inner group within your department	3.27**	2.86
Participate in committees outside the university	3.26**	3.68
Have both M.D. and Ph.D. degrees	3.25	3.14
Have high visibility outside the academic community	3.25	3.41
Counsel students	3.23***	3.91
Teach Inteflex students	3.21	3.32
Negotiate promotion when hired	3.21**	2.71
Teach students in other health professions; e.g., nursing, dentistry, allied health	3.15	3.36
Teach undergraduate students in LS&A	2.97	3.05
Hold a joint appointment	2.91	2.95
Have a heavy formal teaching load	2.76	2.86
Pursue research interests which differ from departmental interests	2.75	2.81
Get unfavorable student evaluations for teaching	2.24*	1.86

PRECIS

CLINICAL PERFORMANCE RATING SCALES

MODERATOR: Charles P. Friedman, Ph.D.
University of North Carolina

THE DEVELOPMENT OF RATING SCALES TO ASSESS THE CLINICAL PERFORMANCE OF MEDICAL STUDENTS

Two rating scales were developed and pilot tested in seven institutions by medical students, supervising residents and faculty. The data collected during the pilot test can be used to improve the reliability and validity of the scales by eliminating items showing a concentration of ratings on one response option, the items which residents and faculty frequently cannot judge and those with low item interrater reliabilities.

INTER-RATER RELIABILITY OF CLINICAL PERFORMANCE RATINGS

The development of rating scales of medical students' clinical skills is reviewed and the development of a fifteen-item behaviorally anchored rating scale is described. The intraclass correlation coefficients for each of the fifteen items and the total scale are presented for ratings made by house officers and attending staff over a twelve-month period.

THE RELATIONSHIP BETWEEN REPORTED ACTIVITIES AND CLINICAL EVALUATIONS OF THIRD-YEAR MEDICAL STUDENTS IN A MEDICINE CLERKSHIP

The correlations between total time spent in twelve medicine clerkship activities and house officer and attending staff evaluations of problem solving and interpersonal skills are computed. Significant positive correlations between (1) time spent in teaching rounds and problem solving, and (2) time spent in work rounds and interpersonal skills were found for both house officers and attending staff evaluations.

202

THE DEVELOPMENT OF RATING SCALES TO ASSESS THE CLINICAL PERFORMANCE OF MEDICAL STUDENTS

Vivian Erviti, Ph.D., Lawrence J. Fabrey, M.S. and Jane V. Bunce
National Board of Medical Examiners

Introduction

Traditionally, evaluation has played an important role in the physician's educational experiences from the time of application for admission to medical school through the time when the physician is ready to assume independent responsibility for patient care. The assessment of physician competency is frequently based on the completion of a specified period of training, performance on written and/or oral examinations, ratings of traits and personality characteristics and ratings of clinical skills. Each of these methods can be effective in assessing some aspects of physician competency.[1,2,3,4]

The evaluation of the clinical performance of medical students is frequently based on ratings of behavior in clinical settings. A rating involves an evaluative summary of past or present experiences in which the rater processes data in a complex and individualistic way to arrive at a judgment. Many factors affect the reliability and validity of the ratings obtained. Raters may be unwilling to spend the time and energy necessary to provide accurate ratings or they may over-identify with the individual being rated and produce biased ratings. Other factors which interfere with the validity of the judgments include the lack of opportunity of the rater to observe the ratee, the covertness and/or ambiguity of the behavior, a lack of a uniform standard to use in interpretation of either the stimulus item or the response option and individual rater biases and idiosyncrasies.[5] Thus, in order to assure that a scale will produce reliable and valid data, it should be developed carefully, pilot tested thoroughly and revised until reliability is adequate and procedures for the selection of raters and for scale administration have been specified. Because of the frequency of use of rating scales, efforts should be made to improve them.[6]

The study described in this report was designed to develop a rating scale for the reliable evaluation of some of the behaviors and skills required of medical students. This tool may be used as a potential criterion of performance for other measurement instruments such as written examinations or as a measure of the competencies of medical students that cannot be assessed by other means.

Methodology

A project advisory committee developed 400 rating scale items based on various role-defining studies.[7] These items were carefully reviewed by members of the project advisory committee and NBME staff to determine whether each item represented

---

Request for reprints: Vivian Erviti, Ph.D., National Board of Medical Examiners, 3930 Chestnut Street, Philadelphia, Pennsylvania 19104.

The research reported in this article was supported in part by the Department of Health, Education, and Welfare, Health Resources Administration, Contract Number 231-76-0067.

a major component of clinical competence and whether medical students might be expected to vary on the item. In addition, a questionnaire was mailed to faculty in four medical schools in the Philadelphia area to determine whether faculty members considered the items appropriate for rating the clinical performance of medical students. Two faculty members in each of the required clerkships were asked to indicate which behavioral statements were appropriate and nineteen out of 40 questionnaires were returned.

Using all data obtained from the surveys and staff reviews, the committee selected 92 items for inclusion in the scales. These items were considered appropriate for students in basic clerkships and those in their most advanced level of a clerkship in primary care disciplines. These students may be in their second, third, or fourth year of medical school, depending upon the program of the student and the curriculum of the participating school.

The project staff randomly assigned 46 items to each of two alternate forms (Form A and Form B). Sample items include: Does not recognize his/her own limitations; Suggests treatment prematurely; Presents cases well. Five response options were provided: Seldom, Sometimes, Usually, Consistently, and Don't Know.

Several medical schools were identified as sites for pilot testing. The NBME staff first contacted the dean's office at each participating institution to obtain the necessary introduction to departments of family practice, internal medicine, obstetrics and gynecology, pediatrics, and surgery. Administrators in the departments agreeing to participate were provided with packets of materials to be used in soliciting student participants. Each medical student who agreed to participate was asked to rate himself/herself and to agree to be rated by two or more supervisors. Each departmental administrator identified those residents and faculty members who had intimate knowledge of a participating student's performance and asked them to rate the student. The administrator was asked to identify at least two resident/faculty raters if it was possible to do so. All participants from a given institution used the same form of the instrument. The number of students contacted and the number who participated are shown in Table 1.

Results

One hundred and two student self-ratings, 52 ratings by residents and 91 ratings by faculty members were completed and returned to the NBME. Three students submitted a self-rating and were rated by four supervisors. Sixteen were rated by three supervisors and an additional 16 by two residents and/or faculty members. A single supervisor rating was received for 39 students and no supervisor ratings were received to correspond to 28 self-ratings.

After data files were established and confidentiality assured, several analyses were performed on the data. The distribution of ratings across the four response options for each item were tabulated for students, residents and staff physicians separately and the means and variances of the items calculated. Most items showed a good distribution across the options and concentration on one or two of the options occurred rarely.

None of the items on Form A showed a concentration of responses on a single response option and only three items on Form B had such a concentration. On Form A, student self-ratings covered three or more options for 39 of the 46 items. Residents used more than two options for 33 items and faculty used three or more on 43 items. The 46 items on Form A had a mean of 3.27 and a standard deviation of .66 on the four-point scale from Seldom to Consistently, with four indicating the most positive behavior. Individual item means ranged from 1.99 to 3.94, and standard deviations ranged from .22 to 1.06. For Form B, students covered three or more options on 43 items, residents covered three or more options on 39 items and staff physicians used them on 34 items. The mean of the item means for Form B was 3.29, ranging from 2.60 to 3.97. The mean standard deviation was .67 with a range of .24 to .97. Thus, all response options were used by some raters at some times for most of the items and the range of item means indicates that raters were not reluctant to use the more negative response options at times.

In general, students were able to provide a rating on more of the 46 items than physicians. Fifty-one percent of the students did not use the Don't Know option and 91% used it less than five times while completing the scale. However, 58% of the residents and 23% of the faculty used it less than five times. The Don't Know response was chosen as many as 27 times by an individual resident and as many as 39 times by a faculty rater. The mean number of times this option was used was significantly different for the two rater groups ($t=5.93$, $p<.001$) as shown in Table 2. The mean number of hours of contact with a student reported by residents and faculty was also significantly different ($t=8.62$, $p<.001$) as shown in Table 2. In addition, there was a significant inverse relationship ($r=-.52$, $p<.001$) between these two variables; that is, physicians who report spending relatively more time with students tend to use the Don't Know response less.

In order to estimate the interrater reliability of the instrument, a Kappa coefficient for each pair of resident and faculty raters available for each medical student was computed.[8] Because persons in different roles may see quite different aspects of the ratee, comparisons were made only between resident-resident and faculty-faculty pairs. As noted above, some students were rated by two, three or four individuals while 67 received an evaluation by only a single supervisor or by none at all. At least two resident or faculty ratings were received for 35 students. However, five of the 35 received ratings from one resident and one faculty member and therefore were not included in the Kappa calculations. Because a few students were rated by three or more physicians of the same class, a total of 40 resident-resident and faculty-faculty calculations could be made. The distribution of Kappa coefficients for the 40 available rater pairs is shown in Table 3. The average coefficient across all rater pairs was .26, a value indicating a "fair" amount of agreement according to the terminology of Landis and Koch.[9] The coefficients ranged from -.01 to .55 with slightly more than half (23 out of 40) falling in the fair or moderate range. As shown in Table 2, there does not appear to be a pattern of reliability or any systematic differences between the reliability of residents and faculty.

In order to look at the effect of eliminating some of the one-category differences in ratings, the four-point scale (Seldom, Sometimes, Usually, Consistently) was dichotomized. The Seldom and Sometimes categories were combined, as were the Usually and Consistently response options. It was felt that there was a greater difference between the second and third responses than between the first and second or the third and fourth. When viewed in this way across all 40 rater pairs, the overall Kappa coefficient was increased to .53 which is considered moderate agreement.

Discussion

Reliable and valid rating scales are difficult to construct because of the many variables that can affect the quality of the data. Raters differ in the level of ratings they assign and also in the spread of their ratings. Interpretation of the items and response options are frequently highly subjective and may vary widely from one rater to another. In addition, some raters are conservative and rarely rate any one very high or very low while others more frequently go toward the extremes.

Although individual differences in ratings were noted in the data collected from the seven institutions, all response options were used for most items. Therefore, the items appear to discriminate appropriately among students. The purpose for preparing two forms of the scale was to reduce the amount of time required of any participant and still collect data on all 92 items. Because no comparison between the two forms was intended, the sample of students being rated on each form was neither random nor matched in any way. Differences between the means and standard deviations for the two forms may be an artifact of the particular cluster of items, the result of the different samples of students being rated, or resident and faculty differences.

The staff had hypothesized that residents would have more extended contact with students than faculty, be more knowledgeable about student performance and therefore provide more reliable ratings. The data support the first two hypotheses as evidenced by significant differences in the number of contact hours reported and the number of Don't Know responses. However, the third hypothesis concerning increased reliability of resident ratings was not supported.

Conclusions and Implications

The development of any evaluative tool is an iterative process. The data obtained in an initial pilot trial can be used to indicate modifications that must then be tested in further trials. The use of multiple data sources and multiple field trials is especially important in developing rating scales because of the large number of sources of error involved in ratings. A review of the Kappa coefficients for the individual rating scale items will enable the project staff to select clusters of items that may improve the overall reliability. It is possible that the elimination of some items may increase reliability. Also, a two-point scale may be more appropriate based on the marked improvement in the Kappa coefficient when two rather than four response options were used.

In addition to the use of reliability coefficients, item elimination may be based on the failure of an item to discriminate among students, the fact that a large number of residents and faculty reported they didn't know about the specified aspect of behavior, the redundancy of items and an editorial review using generally accepted guidelines for rating scale construction.

The data collected in this initial pilot test of the 92 rating scale items will form a beginning of a data pool on individual items. Following revision of the scale and reduction in the number of items on each form, the project staff will collect additional data on another sample of students and repeat the process of data analysis. It is possible that in the future, such a pool could be used to customize scales for various groups of raters, clerkships, institutions, etc.

References

1. Randolph, J.G., Altman, R.P. & Anderson, K.D. Experience with in-service examinations of pediatric surgical trainees. *Journal of Pediatric Surgery*, December 1977, *12*(6), 857-860.

2. Gallagher, R., Donnelly, M., Scalzi, P.M. & Deighton, M. Toward a comprehensive methodology for resident evaluation. *Proceedings of the Sixteenth Annual Conference on Research in Medical Education*, Washington, D.C., November 9-10, 1977.

3. Tudor, J.M. Performance documentation: How to confirm a resident's progress. *Journal of Medical Education*, April 1978, *53*, 337-343.

4. Stillman, P.L. & Sabers, D.L. Using a competency-based program to assess interviewing skills of pediatric house staff. *Journal of Medical Education*, June 1978, *53*, 493-496.

5. Thorndike, R.L. & Hagen, E.P. *Measurement and evaluation in psychology and education* (4th ed.). New York: John Wiley & Sons, 1977.

6. Levine, H.G. & McGuire, C.H. Rating habitual performance in graduate medical education. *Journal of Medical Education*, April 1971, *46*, 306-311.

7. *The Definition of Clinical Competence in Medicine: Performance Dimensions and Rationales for Clinical Skill Areas*. A report of research conducted for the National Board of Medical Examiners by the American Institutes for Research, Palo Alto, CA., April 1980.

8. Fleiss, J.L. Measuring agreement between two judges on the presence or absence of a trait. *Biometrics*, 1975, *31*(3), 651-659.

9. Landis, J.R. & Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics*, March 1977, *33*, 159-174.

Table 1

Number of Students, Residents, and Faculty Participating

School No.	Students Contacted	Students Participating No.	Students Participating %	Resident Raters	Faculty Raters
1	33	8	24%	7	12
2	24	13	54	6	12
3	95	23	24	7	3
4	8	3	38	3	2
5	32	25	78	24	38
6	48	27	56	3	23
7	3	3	100	2	1
	243	102	42%	52	91

Table 2

Comparison of Resident and Faculty Raters

Raters	No. Don't Know Responses $\overline{X}$	No. Don't Know Responses S.D.	No. Hours Contact with Student $\overline{X}$	No. Hours Contact with Student S.D.
Residents	5.58	6.23	36.65	18.98
Faculty	13.58	9.89	11.68	11.52
	$t = 5.93$ ($p < .001$)		$t = 8.62$ ($p < .001$)	

Table 3

Level of Agreement Between Pairs of Resident and Faculty Raters

		Number of Kappa Coefficients in Each Range			
		Form A		Form B	
Kappa	Level of Agreement	Residents	Faculty	Residents	Faculty
< 0	Poor	0	0	0	1
0 - .20	Slight	1	4	7	4
.21 - .40	Fair	3	4	4	6
.41 - .60	Moderate	1	2	1	2
		5	10	12	13

INTER-RATER RELIABILITY OF CLINICAL PERFORMANCE RATINGS

T. E. Dielman, Ph.D., Alan L. Hull, M.S., and Wayne K. Davis, Ph.D.
The University of Michigan

The general phenomenon of inter-observer error as an important source of measurement error has been a subject of investigation in social science and medicine since the 1940's. Fletcher and Oldham (1964) listed a bibliography on observer error which included over seventy papers in the areas of clinical studies, radiology, pathology, and related fields. Landis and Koch (1975a, 1975b) have provided a review of statistical methods in the analysis of inter-rater reliability.

Assessment of the performance of medical students during their clerkship is typically a subjective process involving ratings of complex patterns of behavior. Attempts to quantify such ratings have appeared in the medical education literature during the past twenty-five years.

Diederich (1954) proposed a "critical incidents" technique, derived from studies based on the assessment of Air Force pilots, to assess clinical performance. Cowles and Kubany (1959) adapted Diederich's method and, in interviews with twelve experienced medical faculty members, developed a list of eight characteristics which were considered as most important in preparing for general practice and most easily observed in faculty-student interactions during the third and fourth years of medical school. Those characteristics were identified as: knowledge of medical information; ability to gain and maintain patient's confidence; assumption of responsibility for, and insight into, patient's total problem (medical, social, and emotional); skill in observing, recording, and reporting; skill in developing and verifying hypotheses from patient data; stability under difficult situations; integrity (honesty, ethics, recognition of own limitations) and interest in the profession and in self-improvement. An average of four preceptors rated each of forty-one University of Pittsburgh medical students on these eight scales. The inter-rater reliabilities of the average ratings ranged from -.10 (assumption of responsibility for patient's total problem) to .69 (skill in developing and verifying hypotheses), with a median of .49. The inter-rater reliability of the total of the eight scales was .58.

Cowles (1965) again employed an adaptation of the critical incidents procedure of evaluation and asked medical and non-medical faculty to sort comments made on 190 University of Pittsburgh medical students into groups. Ten groupings resulted from this process which were identified as: knowledge, rapport with patients, assumption of responsibility, accuracy and thoroughness of observation, diagnostic skill, recognition of limitations, interest/motivation, general intelligence, maturity/poise, and general excellence in medicine. The inter-rater reliabilities were reported as "high" although no data were presented.

Gough, et al. (1964) analyzed faculty ratings on ten variables in a study of eighty-one medical students during their third- and fourth-year clerkships. The variables and their inter-rater reliabilities for the two years were: clinical judgment [.72, .73]; diagnostic skill [.71, .76];

Send requests for reprints to T. E. Dielman, Ph.D., Office of Educational Resources and Research, The University of Michigan Medical School, G1210 Towsley Center, Ann Arbor, Michigan 48109.

mastery of basic sciences [.60, .75]; effectiveness of manner with patients [.57, .78]; sense of responsibility [.61, .74]; independence [.49, .77]; identification with medicine [.42, .65]; general promise for internship and residency training [.66, .82]; and success potential [.69, .80].

Printen, et al. (1973) have reported a clinical performance evaluation of junior medical students. The rating variables employed in that study were extracted from those used by the University of Illinois College of Medicine. The estimated correlation between clinical performance rating by two single raters averaged across the ten rating scales was .64, while the probability of a student receiving the same score from two separate groups of six raters was calculated to be .92.

The rating scales developed for the current study were based on instruments employed by various medical schools, most of which were influenced by the research which has been cited. On the basis of the literature review and discussions with the faculty, departmental committees, and students of The University of Michigan Medical School, a decision was made to develop a set of behaviorally anchored scales in order to provide specific feedback to the students and to provide the rater with a behavioral rather than a vague conceptual definition of the trait under consideration. The purpose of the present study was to provide evidence concerning the inter-rater reliability of those scales.

Method

The clinical evaluation form used in the current study consists of fifteen behaviorally based performance scales (See Figure 1). The categories represent a range of clinical and professional skills. The form provides four behaviorally defined categories for rating each of the fifteen performance scales. A space for "Not Observed" is also provided on each scale. The behaviorally defined categories were developed in order to provide the students with more specific feedback than would have been provided by adjectives such as "Outstanding" or "Poor."

The data base for the inter-rater reliability analysis included 1,908 ratings of 389 students by attending staff and 1,749 ratings of 355 students by house officers. The ratings were gathered over a twelve-month period from July, 1977, to July, 1978, on third- and fourth-year medical students. The junior clerkship year is divided into twelve four-week periods; the senior year into ten. Students were rated during each of the periods by two or more house officers and/or attending staff members. The inter-rater reliability for each of the fifteen rating scales was computed separately for house officers and attending staff. Inter-rater reliabilities for the fifteen scales were calculated by an ordinal data extension of the categorical data computation of intraclass correlation coefficients as discussed by Landis and Koch (1977).

The intraclass correlation coefficient is basically a ratio of the within subject variability to the total variance, which is composed of variance due to subjects, observers, and error. The within subject intraclass correlation coefficient where d observers are assumed to be a random sample of potential observers is given by:

$$\rho = \frac{\hat{\sigma}_s^2}{\hat{\sigma}_s^2 + \hat{\sigma}_d^2 + \hat{\sigma}_e^2}$$

where,

$\hat{\sigma}_s^2$ is the estimate of the variance due to differences in ratees over all observers, or the "among subjects effects,"

$\hat{\sigma}_d^2$ is the estimate of the variance due to differences in the average judgment of observers, or the "among observers effects," and

$\hat{\sigma}_e^2$ is the estimate of the intrinsic measurement error variance component.

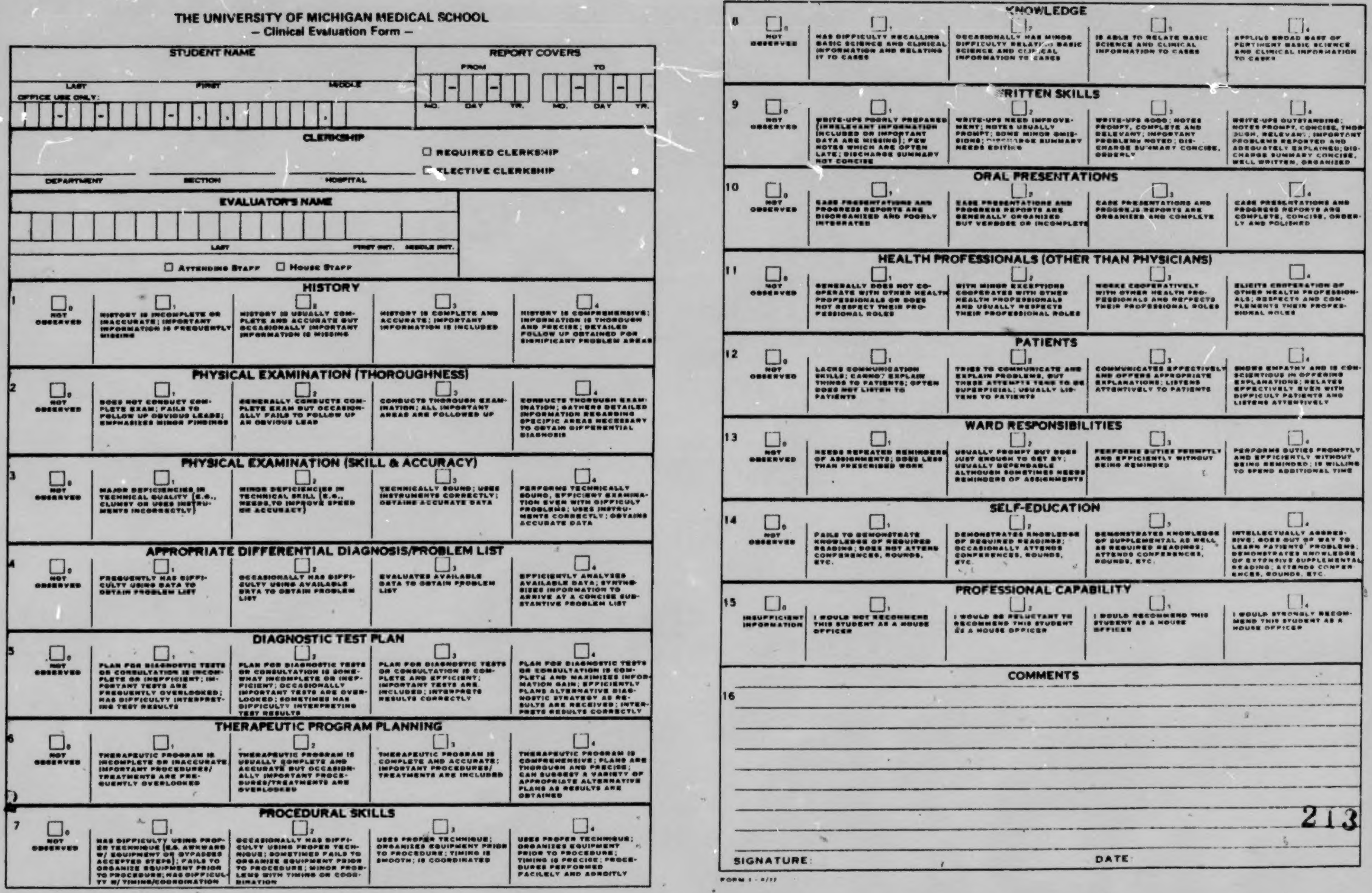
Results

The inter-rater reliability coefficients for each item and the total scale are presented separately for house officers and attending faculty in Table 1. The inter-rater agreement on the fifteen individual items for the attending faculty ranged from .22 to .37, with a median coefficient of .29. The attending faculty inter-rater agreement for the fifteen items summed was .40. For house officers, the inter-rater agreement for the fifteen single items ranged from .30 to .51, with a median of .36. The house officer agreement was .61 for the summation of the fifteen items.

Discussion

The inter-rater reliability on the single items was somewhat higher for house officers than for attending faculty. One explanation which has been proposed for this result is that house officers spend more time observing the students than do the attending faculty, and thus have more data on which to base their ratings. The strength of inter-rater agreement, as judged along the admittedly arbitrary benchmarks employed by Landis and Koch (1977), is fair to moderate for both groups of raters. Agreement increases when the individual ratings are summed, and it is recommended that such combined scores be employed for the purpose of summative evaluation. Combining the scales in such a fashion is of little use for formative evaluation purposes, however. When the scales are used for formative purposes in providing student feedback, it is recommended that individual ratings be provided and discrepancies among ratings rendered by different raters be discussed. Another possibility when reporting the results of ratings on individual items is to average ratings made by several observers, thus eliminating the error due to observer variation. This procedure can increase the reliability of ratings on single items considerably, as Printen, et al. (1973) have shown. Low inter-rater reliability seems to be the norm when dealing with independent subjective judgments, and ratings of medical student performance are no exception. Faculty members should keep this in mind when evaluating students and strive to collect several independent judgments.

Figure 1. The Clinical Evaluation Form



THE UNIVERSITY OF MICHIGAN MEDICAL SCHOOL
– Clinical Evaluation Form –

STUDENT NAME LAST FIRST MIDDLE OFFICE USE ONLY:	REPORT COVERS FROM MO. DAY YR. TO MO. DAY YR.
CLERKSHIP DEPARTMENT SECTION HOSPITAL	☐ REQUIRED CLERKSHIP ☐ ELECTIVE CLERKSHIP
EVALUATOR'S NAME LAST FIRST INIT. MIDDLE INIT. ☐ ATTENDING STAFF ☐ HOUSE STAFF	

HISTORY

1	☐ 0 NOT OBSERVED	☐ 1 HISTORY IS INCOMPLETE OR INACCURATE; IMPORTANT INFORMATION IS FREQUENTLY MISSING	☐ 2 HISTORY IS USUALLY COMPLETE AND ACCURATE BUT OCCASIONALLY IMPORTANT INFORMATION IS MISSING	☐ 3 HISTORY IS COMPLETE AND ACCURATE; IMPORTANT INFORMATION IS INCLUDED	☐ 4 HISTORY IS COMPREHENSIVE; INFORMATION IS THOROUGH AND PRECISE; DETAILED FOLLOW UP OBTAINED FOR SIGNIFICANT PROBLEM AREAS

PHYSICAL EXAMINATION (THOROUGHNESS)

2	☐ 0 NOT OBSERVED	☐ 1 DOES NOT CONDUCT COMPLETE EXAM; FAILS TO FOLLOW UP OBVIOUS LEADS; EMPHASIZES MINOR FINDINGS	☐ 2 GENERALLY CONDUCTS COMPLETE EXAM BUT OCCASIONALLY FAILS TO FOLLOW UP AN OBVIOUS LEAD	☐ 3 CONDUCTS THOROUGH EXAMINATION; ALL IMPORTANT AREAS ARE FOLLOWED UP	☐ 4 CONDUCTS THOROUGH EXAMINATION; GATHERS DETAILED INFORMATION REGARDING SPECIFIC AREAS NECESSARY TO OBTAIN DIFFERENTIAL DIAGNOSIS

PHYSICAL EXAMINATION (SKILL & ACCURACY)

3	☐ 0 NOT OBSERVED	☐ 1 MAJOR DEFICIENCIES IN TECHNICAL QUALITY (E.G., CLUMSY OR USES INSTRUMENTS INCORRECTLY)	☐ 2 MINOR DEFICIENCIES IN TECHNICAL SKILL (E.G., NEEDS TO IMPROVE SPEED OR ACCURACY)	☐ 3 TECHNICALLY SOUND; USES INSTRUMENTS CORRECTLY; OBTAINS ACCURATE DATA	☐ 4 PERFORMS TECHNICALLY SOUND, EFFICIENT EXAMINATION EVEN WITH DIFFICULT PROBLEMS; USES INSTRUMENTS CORRECTLY; OBTAINS ACCURATE DATA

APPROPRIATE DIFFERENTIAL DIAGNOSIS/PROBLEM LIST

4	☐ 0 NOT OBSERVED	☐ 1 FREQUENTLY HAS DIFFICULTY USING DATA TO OBTAIN PROBLEM LIST	☐ 2 OCCASIONALLY HAS DIFFICULTY USING AVAILABLE DATA TO OBTAIN PROBLEM LIST	☐ 3 EVALUATES AVAILABLE DATA TO OBTAIN PROBLEM LIST	☐ 4 EFFICIENTLY ANALYZES AVAILABLE DATA; SYNTHESIZES INFORMATION TO ARRIVE AT A CONCISE SUBSTANTIVE PROBLEM LIST

DIAGNOSTIC TEST PLAN

5	☐ 0 NOT OBSERVED	☐ 1 PLAN FOR DIAGNOSTIC TESTS OR CONSULTATION IS INCOMPLETE OR INEFFICIENT; IMPORTANT TESTS ARE FREQUENTLY OVERLOOKED; HAS DIFFICULTY INTERPRETING TEST RESULTS	☐ 2 PLAN FOR DIAGNOSTIC TESTS OR CONSULTATION IS SOMEWHAT INCOMPLETE OR INEFFICIENT; OCCASIONALLY IMPORTANT TESTS ARE OVERLOOKED; SOMETIMES HAS DIFFICULTY INTERPRETING TEST RESULTS	☐ 3 PLAN FOR DIAGNOSTIC TESTS OR CONSULTATION IS COMPLETE AND EFFICIENT; IMPORTANT TESTS ARE INCLUDED; INTERPRETS RESULTS CORRECTLY	☐ 4 PLAN FOR DIAGNOSTIC TESTS OR CONSULTATION IS COMPLETE AND MAXIMIZES INFORMATION GAIN; EFFICIENTLY PLANS ALTERNATIVE DIAGNOSTIC STRATEGY AS RESULTS ARE RECEIVED; INTERPRETS RESULTS CORRECTLY

THERAPEUTIC PROGRAM PLANNING

6	☐ 0 NOT OBSERVED	☐ 1 THERAPEUTIC PROGRAM IS INCOMPLETE OR INACCURATE IMPORTANT PROCEDURES/TREATMENTS ARE FREQUENTLY OVERLOOKED	☐ 2 THERAPEUTIC PROGRAM IS USUALLY COMPLETE AND ACCURATE BUT OCCASIONALLY IMPORTANT PROCEDURES/TREATMENTS ARE OVERLOOKED	☐ 3 THERAPEUTIC PROGRAM IS COMPLETE AND ACCURATE; IMPORTANT PROCEDURES/TREATMENTS ARE INCLUDED	☐ 4 THERAPEUTIC PROGRAM IS COMPREHENSIVE; PLANS ARE THOROUGH AND PRECISE; CAN SUGGEST A VARIETY OF APPROPRIATE ALTERNATIVE PLANS AS RESULTS ARE OBTAINED

PROCEDURAL SKILLS

7	☐ 0 NOT OBSERVED	☐ 1 HAS DIFFICULTY USING PROPER TECHNIQUE (E.G. AWKWARD W/ EQUIPMENT OR BYPASSES ACCEPTED STEPS); FAILS TO ORGANIZE EQUIPMENT PRIOR TO PROCEDURE; HAS DIFFICULTY W/ TIMING/COORDINATION	☐ 2 OCCASIONALLY HAS DIFFICULTY USING PROPER TECHNIQUE; SOMETIMES FAILS TO ORGANIZE EQUIPMENT PRIOR TO PROCEDURE; MINOR PROBLEMS WITH TIMING OR COORDINATION	☐ 3 USES PROPER TECHNIQUE; ORGANIZES EQUIPMENT PRIOR TO PROCEDURE; TIMING IS SMOOTH; IS COORDINATED	☐ 4 USES PROPER TECHNIQUE; ORGANIZES EQUIPMENT PRIOR TO PROCEDURE; TIMING IS PRECISE; PROCEDURES PERFORMED FACILELY AND ADROITLY

KNOWLEDGE

8	☐ 0 NOT OBSERVED	☐ 1 HAS DIFFICULTY RECALLING BASIC SCIENCE AND CLINICAL INFORMATION AND RELATING IT TO CASES	☐ 2 OCCASIONALLY HAS MINOR DIFFICULTY RELATING BASIC SCIENCE AND CLINICAL INFORMATION TO CASES	☐ 3 IS ABLE TO RELATE BASIC SCIENCE AND CLINICAL INFORMATION TO CASES	☐ 4 APPLIES BROAD BASE OF PERTINENT BASIC SCIENCE AND CLINICAL INFORMATION TO CASES

WRITTEN SKILLS

9	☐ 0 NOT OBSERVED	☐ 1 WRITE-UPS POORLY PREPARED (IRRELEVANT INFORMATION INCLUDED OR IMPORTANT DATA ARE MISSING); FEW NOTES WHICH ARE OFTEN LATE; DISCHARGE SUMMARY NOT CONCISE	☐ 2 WRITE-UPS NEED IMPROVEMENT; NOTES USUALLY PROMPT; SOME MINOR OMISSIONS; DISCHARGE SUMMARY NEEDS EDITING	☐ 3 WRITE-UPS GOOD; NOTES PROMPT, COMPLETE AND RELEVANT; IMPORTANT PROBLEMS NOTED; DISCHARGE SUMMARY CONCISE, ORDERLY	☐ 4 WRITE-UPS OUTSTANDING; NOTES PROMPT, CONCISE, THOROUGH, RELEVANT; IMPORTANT PROBLEMS REPORTED AND ADEQUATELY EXPLAINED; DISCHARGE SUMMARY CONCISE, WELL WRITTEN, ORGANIZED

ORAL PRESENTATIONS

10	☐ 0 NOT OBSERVED	☐ 1 CASE PRESENTATIONS AND PROGRESS REPORTS ARE DISORGANIZED AND POORLY INTEGRATED	☐ 2 CASE PRESENTATIONS AND PROGRESS REPORTS ARE GENERALLY ORGANIZED BUT VERBOSE OR INCOMPLETE	☐ 3 CASE PRESENTATIONS AND PROGRESS REPORTS ARE ORGANIZED AND COMPLETE	☐ 4 CASE PRESENTATIONS AND PROGRESS REPORTS ARE COMPLETE, CONCISE, ORDERLY AND POLISHED

HEALTH PROFESSIONALS (OTHER THAN PHYSICIANS)

11	☐ 0 NOT OBSERVED	☐ 1 GENERALLY DOES NOT COOPERATE WITH OTHER HEALTH PROFESSIONALS OR DOES NOT RESPECT THEIR PROFESSIONAL ROLES	☐ 2 WITH MINOR EXCEPTIONS COOPERATES WITH OTHER HEALTH PROFESSIONALS AND USUALLY RESPECTS THEIR PROFESSIONAL ROLES	☐ 3 WORKS COOPERATIVELY WITH OTHER HEALTH PROFESSIONALS AND RESPECTS THEIR PROFESSIONAL ROLES	☐ 4 ELICITS COOPERATION OF OTHER HEALTH PROFESSIONALS; RESPECTS AND COMPLEMENTS THEIR PROFESSIONAL ROLES

PATIENTS

12	☐ 0 NOT OBSERVED	☐ 1 LACKS COMMUNICATION SKILLS; CANNOT EXPLAIN THINGS TO PATIENTS; OFTEN DOES NOT LISTEN TO PATIENTS	☐ 2 TRIES TO COMMUNICATE AND EXPLAIN PROBLEMS, BUT THESE ATTEMPTS TEND TO BE SUPERFICIAL; USUALLY LISTENS TO PATIENTS	☐ 3 COMMUNICATES EFFECTIVELY AND OFFERS APPROPRIATE EXPLANATIONS; LISTENS ATTENTIVELY TO PATIENTS	☐ 4 SHOWS EMPATHY AND IS CONSCIENTIOUS IN OFFERING EXPLANATIONS; RELATES EFFECTIVELY EVEN WITH DIFFICULT PATIENTS AND LISTENS ATTENTIVELY

WARD RESPONSIBILITIES

13	☐ 0 NOT OBSERVED	☐ 1 NEEDS REPEATED REMINDERS OF ASSIGNMENTS; DOES LESS THAN PRESCRIBED WORK	☐ 2 USUALLY PROMPT BUT DOES JUST ENOUGH TO GET BY; USUALLY DEPENDABLE ALTHOUGH SOMETIMES NEEDS REMINDERS OF ASSIGNMENTS	☐ 3 PERFORMS DUTIES PROMPTLY AND EFFICIENTLY WITHOUT BEING REMINDED	☐ 4 PERFORMS DUTIES PROMPTLY AND EFFICIENTLY WITHOUT BEING REMINDED; IS WILLING TO SPEND ADDITIONAL TIME

SELF-EDUCATION

14	☐ 0 NOT OBSERVED	☐ 1 FAILS TO DEMONSTRATE KNOWLEDGE OF REQUIRED READING; DOES NOT ATTEND CONFERENCES, ROUNDS, ETC.	☐ 2 DEMONSTRATES KNOWLEDGE OF REQUIRED READINGS; OCCASIONALLY ATTENDS CONFERENCES, ROUNDS, ETC.	☐ 3 DEMONSTRATES KNOWLEDGE OF SUPPLEMENTAL AS WELL AS REQUIRED READINGS; ATTENDS CONFERENCES, ROUNDS, ETC.	☐ 4 INTELLECTUALLY AGGRESSIVE; GOES OUT OF WAY TO LEARN PATIENTS' PROBLEMS; DEMONSTRATES KNOWLEDGE OF EXTENSIVE SUPPLEMENTAL READING; ATTENDS CONFERENCES, ROUNDS, ETC.

PROFESSIONAL CAPABILITY

15	☐ 0 INSUFFICIENT INFORMATION	☐ 1 I WOULD NOT RECOMMEND THIS STUDENT AS A HOUSE OFFICER	☐ 2 I WOULD BE RELUCTANT TO RECOMMEND THIS STUDENT AS A HOUSE OFFICER	☐ 3 I WOULD RECOMMEND THIS STUDENT AS A HOUSE OFFICER	☐ 4 I WOULD STRONGLY RECOMMEND THIS STUDENT AS A HOUSE OFFICER

COMMENTS

16

SIGNATURE: DATE:

FORM 1 - 8/77

Table 1. INTER-RATER RELIABILITIES FOR THE FIFTEEN SINGLE ITEMS AND THE TOTAL SCALE AS RATED BY HOUSE OFFICERS AND ATTENDING FACULTY (ALL PERIODS)

Inter-rater Reliabilities	House Officers (N=1,749)*	Attending Faculty (N=1,908)*
Item		
1. History	.34	.34
2. Physical Examination (Thoroughness)	.33	.28
3. Physical Examination (Skill & Accuracy)	.31	.22
4. Differential Diagnosis	.39	.30
5. Diagnostic Test Plan	.37	.33
6. Therapeutic Program	.39	.27
7. Procedural Skills	.34	.23
8. Knowledge	.42	.27
9. Written Skills	.35	.28
10. Oral Presentations	.36	.35
11. Relationships With Other Professionals	.30	.29
12. Relationships With Patients	.30	.31
13. Ward Responsibilities	.42	.37
14. Self-Education	.38	.25
15. Professional Capability	.51	.36
TOTAL SCALE	.61	.40

*NOTE: N's refer to the total number of ratings over the twelve periods.

REFERENCES

Cowles, J. T. A critical-comments approach to the rating of medical students' clinical performance. Journal of Medical Education, 1965, 40, 188-198.

Cowles, J. T. and Kubany, A. J. Improving the measurement of clinical performance of medical students. Journal of Medical Education, 1959, 15, 139-142.

Diederich, P. B. The critical incidents technique applied to medical education. ETS Research Memorandum 54-9. Princeton: Educational Testing Service, 1954.

Fletcher, C. M. and Oldham, P. D. Diagnosis in group research. In L. J. Witts (Ed.), Medical Surveys and Clinical Trials. London: Oxford University Press, 1964.

Gough, H. G., Hall, W. B., and Harris, R. E. Evaluation of performance in medical training. Journal of Medical Education, 1964, 39, 679-692.

Landis, J. R. and Koch, G. G. The measurement of observer agreement for categorical data. Biometrics, 1977, 33, 159-174.

Landis, J. R. and Koch, G. G. A review of statistical methods in the analysis of data arising from observer reliability studies (Part I). Statistica Neerlandica, 1975, 29, 101-123. (a)

Landis, J. R. and Koch, G. G. A review of statistical methods in the analysis of data arising from observer reliability studies (Part II). Statistica Neerlandica, 1975, 29, 151-161. (b)

Printen, K. J., Chappell, W., and Whitney, D. R. Clinical performance evaluation of junior medical students. Journal of Medical Education, 1973, 48, 343-348.

THE RELATIONSHIP BETWEEN REPORTED ACTIVITIES AND CLINICAL EVALUATIONS OF THIRD-YEAR MEDICAL STUDENTS IN A MEDICINE CLERKSHIP

Alan L. Hull, M.S., Judith G. Calhoun, Ph.D., and Wayne K. Davis, Ph.D.
The University of Michigan Medical School

Background

Clinical clerkship performance of medical students has traditionally been based on two methods of evaluation: (1) examination, and (2) observations by attending staff and house officers. Although a review of the literature on medical education reveals little relationship between cognitive knowledge scores and clinical performance scores and ultimate success in delivering medical care,[1,2] tests measuring intellectual skills remain a major factor in assessing the numerous factors related to clerkship experiences and achievement. This tendency to center clerkship evaluation around recall of information is due to the ease with which this dimension can be measured. Other student clerkship performance behaviors and skills are much more difficult to quantify for reliable assessment within the context of the clerkship setting.

Several studies have been reported in the literature which relate student allocation of time to achievement scores. In their study of how four classes of medical students spend their time at the University of Illinois, Fisher and Cotsonas[3] found no significant correlation between student assessment scores and the time spent in formally scheduled activities or with time spent in home study or other study. Only two variables in their study showed significant correlations with achievement. For the freshman class, time spent in travel to and from the university was negatively correlated with achievement. For the sophomores, leisure and recreation was positively correlated with achievement.

Levitsky, et al.[4] found the important clerkship evaluation characteristics such as student motivation, attitude and behavioral identification with professorial models too difficult to measure and define objectively. Therefore, cognitive knowledge and problem-solving factors, as measured by a university clinical comprehensive examination and Part II of the National Board of Medical Examiners test, were used to determine medical student achievement in university and non-university clerkship settings. While the examination scores revealed similarities in cognitive knowledge between the two types of hospitals, they had a low correlation with clinical performance scores. The authors suggested that these instruments were perhaps measuring different aspects of clerkship competence. Therefore, they proposed that unless a systematic approach was made to operationally define clinical performance, the final clerkship grade or indicator of performance would continue to be based on the predominately unstructured, informal personal encounters that usually transpire between faculty and students during the clerkship. Hence, inter-rater agreement of critical clerkship components would be low in this dimension of clinical assessment.

Reprints: Alan L. Hull, Office of Educational Resources and Research,
The University of Michigan Medical School, G1113 Towsley Center,
Ann Arbor, MI 48109.

Linn and Zeppa[5] reported that student clerkship evaluations were often primarily based upon examination scores due to the nonspecificity and immeasurability of other non-cognitive clinical skills. Most other studies found in the literature report the relationships between test results (e.g., NBME, shelf exams, written exams) and clinical performance. However, few studies have addressed time allocation and clinical assessment. Therefore the purpose of this paper is to study the relationship between a student's allocation of time among twelve activities in a medicine clerkship and the evaluation of the student's clinical performance by faculty and house officers.

The data used in this study were collected from two data sources, a structured clerkship performance evaluation form and self-reported activities of students. Two instruments were used to collect the data for these data sources. The first was the Clinical Evaluation Form (CEF) developed at The University of Michigan Medical School. The instrument consists of fifteen four-point, behaviorally anchored items. The instrument has been demonstrated to be reliable, encompassing two dimensions of a student's experience: problem solving and interpersonal skills. (See Dielman, et al.[6] for a discussion of the instrument's reliability and factor structure.) The second data source consisted of time logs completed by 149 students describing their activities during a typical week of their medicine clerkship. The data collected during the log study have been analyzed to determine if differences in activities exist either among various hospital locations or during the academic year. (See Calhoun, et al.[7] for a description of the log study and results of the study.)

The medicine clerkship is twelve weeks in duration and is offered four times during the junior year. The sites for the clerkship include the University hospital, the Veterans Administration Hospital, a county and a private hospital setting.

Method

Components of two data files were combined to develop the research data base for this study. First, one evaluation form completed by a house officer and another by an attending staff was randomly selected for each student during the last month of their medicine clerkship. The last month of the clerkship was selected since it coincides with the data collection time schedule in the log study. Since the CEF system is voluntary, not all students had forms completed by either one or both of the evaluator types. In addition, the random selection procedure was modified to include only those CEF evaluations which had responses to at least one-half of the items in each of the two scales. These criteria resulted in the selection of 263 CEF evaluations for 140 different students. Mean scores were calculated for the problem-solving and interpersonal skills subscales for all non-missing items in each subscale.

Secondly, the time data for each log were aggregated by type of activity for the 149 students who participated in the log study. Since there were significant differences between the time spent in these activities on week days and weekends, only week day times were included in the aggregates. Table 1 displays the twelve activities.

Table 1

Clerkship Activities Included in Log Study
Initial History and Physical
Write-up
Chart Work (notes, reviews, etc.)
Interaction with Hospitalized Inpatient
Procedure (IV, Lp, etc.)
Lab Work
Work Rounds
Teaching Rounds
Teaching Conference
Discussion of Medical Topic (NOT with patient)
Seminar/Lecture
Individual Study or Research

Since no one was required to participate in either the log study or the CEF system, it was possible to have a substantial amount of missing data. Matching the 149 students in the log study with the 140 different students who received house officer and/or attending staff CEF evaluations resulted in 100 student matches for attending staff evaluation-time data, and 115 complete cases for house officer evaluation-time data. Table 2 displays the distribution of the students among the four hospitals included in the study. Separate analyses were not conducted for each of the hospitals due to the large variation in numbers of students among the four hospital sites.

Table 2

Distribution of Students Among Hospitals

	Hospital				
Evaluator Type	1	2	3	4	Total
Attending	13	58	23	6	100
House Officer	13	67	25	10	115

The statistical technique used to study the relationship between distribution of time and student evaluations was to compute product-moment correlations between the two sets of data. Correlations were calculated separately for students who received evaluations from attending staff and house officer staff. These correlations were computed between the CEF problem solving and interpersonal subscales and time spent in each of the twelve clerkship activities.

Results

The most consistent findings for students rated by attending staff and by house officers were (1) that time spent in teaching rounds was significantly correlated with problem solving ratings, and (2) time spent in work rounds (i.e., students caring for their own patients) was significantly correlated with ratings of interpersonal skills. The only other significant correlations were between attending staff ratings of interpersonal skills and time spent in interactions with hospitalized patients, and a negative correlation between attending staff rating of problem solving and amount of time spent in teaching conferences. Table 3 displays the correlations which were calculated between the CEF evaluations and the twelve clerkship activities as well as the three clerkship activity groups.

Table 3

Correlations Between CEF Evaluations and Time Activities

	CEF Evaluations Subscale			
	Knowledge		Interpersonal Skill	
Activity	Attending Staff (N=100)	House Officer (N=115)	Attending Staff (N=100)	House Officer (N=115)
History	-.07	.04	.12	-.04
Write-up	.06	.14	.10	-.01
Chart work	.11	.14	-.05	-.02
Patient Interaction	.11	.17	.22*	.11
Procedure	-.02	.04	.12	.07
Lab work	-.04	.02	-.07	-.03
Work rounds	.10	.13	.21*	.20*
Teaching rounds	.23*	.22*	.14	.09
Teaching conference	-.19*	-.11	-.08	-.06
Discussion	.02	-.09	-.11	.02
Seminar	.04	.06	.05	.15
Individual Study	.16	.14	.17	.18

* = p<.05

Discussion

The results of this study clearly indicate that students who spend more time in work and teaching rounds will benefit from higher faculty and house officer evaluations on both the problem solving and interpersonal skills dimensions. The sensitivity of the time allocation to differentiate between work rounds and teaching rounds is enlightening. Students were made aware of the difference between the two types of rounds when the log data were collected which reduces the possible error due to misclassifying an activity. Teaching rounds require the students to demonstrate their knowledge of medicine and ability to recall and apply information learned during their basic science training. Work rounds require the student to develop a working relationship with his or her patient, requiring the student to develop a good "bedside manner." This phenomenon is also reflected in the significant correlation between attending staff ratings of interpersonal skills and amount of time spent in patient interactions. The negative correlation between time spent in teaching conference and attending staff evaluations of problem solving is striking. Teaching conferences are one of the most structured components of the medicine clerkship curriculum, and it is possible that students spending more time in this activity have more of an opportunity to display a lack of knowledge to the attending staff.

The implication of this study is that both faculty and house officers appear to develop their opinions of student ability during work and teaching rounds. This study is limited in that the activity variable studied is total time; there is no indication of the medical student's role or even the content of what is done in each of the activities. However, the results tend to contradict the studies by Fisher and Cotsonas[3] and Linn and Zeppa[5] because of the significant correlation between time spent in round activities and an evaluation of a student's performance on the problem solving and interpersonal skills subscales. Levitsky, et al. concluded that evaluation of clinical performance is based upon informal and unstructured personal encounters in the absence of formal criteria is not applicable in this study. The CEF includes fifteen items with explicit performance criteria with which to evaluate the student. Finally, the study demonstrates that faculty and house officers depend on the traditional clinical practice of "going on rounds" for their evaluation of student performance.

REFERENCES

1. Schwartz, G. F. and Gonnella, J. S. The Surgical Clerkship: Assessment of Clinical Competence, J. Surg. Res. 14:373, 1973.

2. Wingard, J. R. and Williamson, J. W. Grades as Predictors of Physicians Career Performance: An Evaluative Literature Review, J. Med. Ed., 48:311-322, 1973.

3. Fisher, L. A. and Cotsonas, N. J. A Time Study of Student Activities, J. Med. Ed., 40:125-131, 1965.

4. Levitsky, S., Nerenberg, R. L., Bombeck, M. D. and Lylus, L. M. Are There Educational Advantages to a Surgical Clerkship in a University Hospital?, Surgery, 75:908-914, 1974.

5. Linn, B. S. and Zeppa, R. Measuring Performance in the Surgical Clerkship, J. Med. Ed., 49:601-604, 1974.

6. Dielman, T. E., Hull, A. L., and Davis, W. K. Properties of Clinical Performance Ratings. Technical Memorandum 79.2, Office of Educational Resources and Research, The University of Michigan Medical School, Ann Arbor, 1979.

7. Calhoun, J. G., Hull, A. L., and Davis, W. K. A Time Allocation Study in the Medicine Clerkship. Technical Memorandum 79.4, Office of Educational Resources and Research, The University of Michigan Medical School, Ann Arbor, 1979.

PRECIS

MEANING, COMMUNICATION, AND MEMORY

MODERATOR: Howard L. Stone, Ph.D.
University of Wisconsin

THE COMMUNICATION OF UNCERTAINTY

The paper describes a study in which novices and experts were asked to assign a numerical probability to words commonly used to communicate uncertainty. There was a general lack of agreement for all descriptors, unrelated to educational level.

PHYSICIAN-PATIENT COMMUNICATION: INTERPRETATIONS OF NON-TECHNICAL PHRASES

This study investigated the interpretation of ten non-technical phrases occurring in physician-patient communication. Results indicated multiple interpretations of these non-technical phrases within the patient group, within the physician group, and between the patient and physician groups.

CLINICAL EXPERIENCE AND THE STRUCTURE OF MEMORY

The paper describes a new measure intended to assess factors related to clinical memory. Construct validity was demonstrated using samples of physicians, residents, and undergraduate students.

THE COMMUNICATION OF UNCERTAINTY

G.D. Bryant and G.R. Norman
Flinders Medical Centre and McMaster University

Medical opinions frequently contain verbal probability estimates. These have received little attention either because they are believed to be unambiguous or else because any imprecision and vagueness is believed only appropriate in view of the subjective nature of opinions. Decision analysts argue that uncertainty should be explicitly quantified using a numerical scale and have developed several techniques for this (1). This study examines the verbal scale and whether there is any potential for misunderstanding between health professionals who use it.

SUBJECTS AND METHODS

The thirty-one subjects were associated with one hospital in the Hamilton area. There were four pathologists, four radiologists, four respirologists, four physicians, five residents, five clinical clerks, and five Phase II McMaster medical students.

The expressions were chosen from radiology and laboratory reports and physicians' notes. Each subject was shown the relationship between probability and odds and then asked to mark on the form the likelihood of the presence of a disease when each expression was used. No specific disease was mentioned and no time limit was set. The test was always supervised by the same investigator and usually took less than ten minutes. Not less than two days later each subject repeated the test giving a total of 62 forms for analysis.

RESULTS

Figure I shows the mean, standard deviation, and range for each expression. It is apparent that there is little consistency between subjects, with no phrases having ranges less than half the scale. A number of phrases have extreme ranges; for example the phrase "odds on" was assigned probability values from .00 to .95. Closer inspection of the distributions showed that for two descriptors, "effectively excludes" and "never", this extreme range was apparently due to a misinterpretation of the instructions, as the distributions were biomodal.

An analysis-of-variance showed that for five phrases, there was a significant difference in the mean probabilities assigned by various groups. Generally speaking, when differences occurred, the interpretations of the pathologists and radiologists were more extreme than the other groups.

Reprint requests: G.R. Norman, Ph.D., Department of Family Medicine,
McMaster University, 1200 Main Street West,
Hamilton, Ontario, Canada, L8S 4J9

To test whether the large observed differences were due to the inexperience of the student groups, Levine's test of equal variances was used. Eight descriptors were associated with statistically significant differences in within-group variance. In two cases, the laboratory physicians had lower variance than the other groups; the student groups had lower variation and therefore greater agreement for four phrases, and in only one phrase, "pathognomic", did the student group show markedly less agreement than the remaining samples.

DISCUSSION

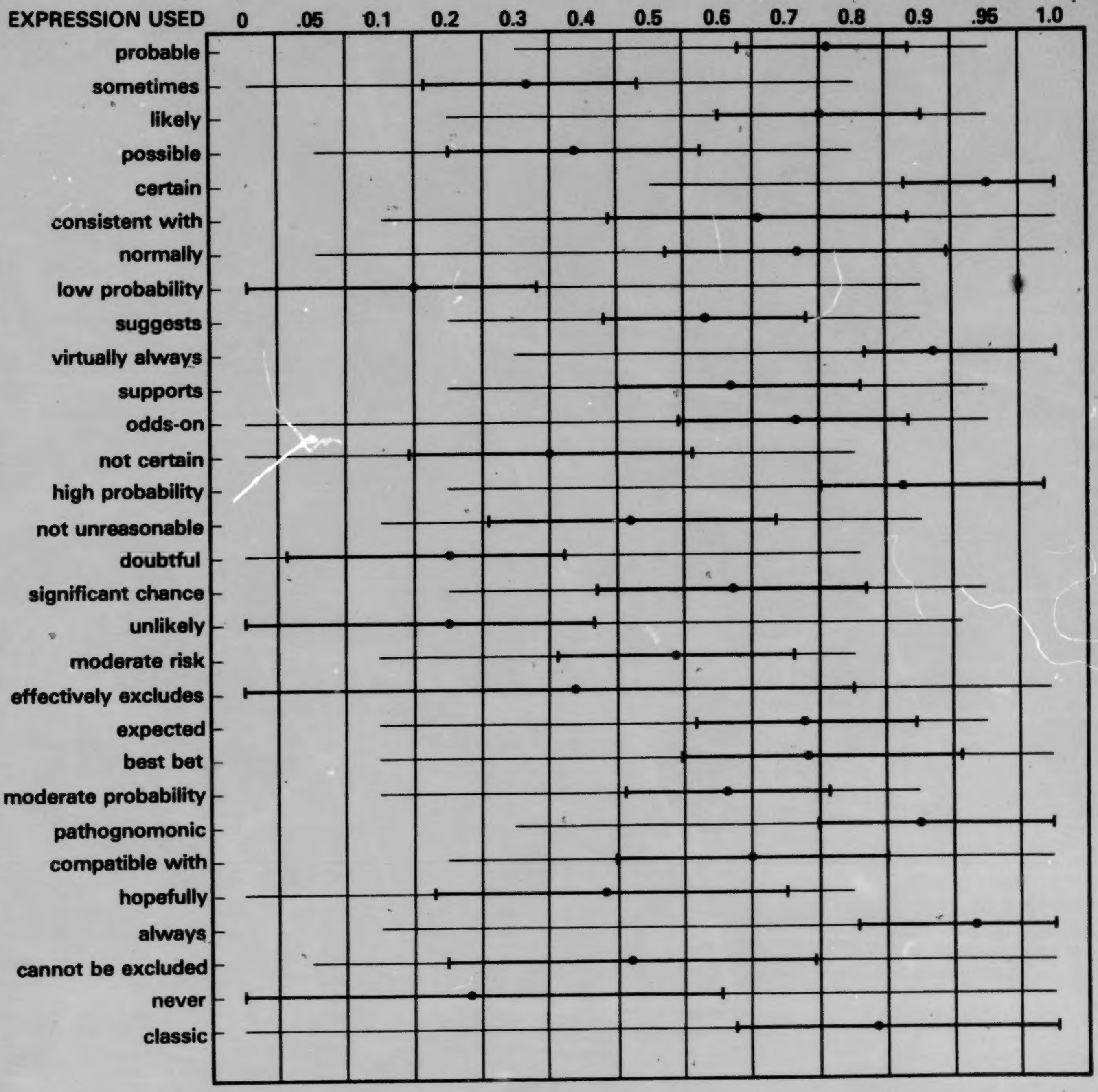
This study shows verbal probability estimates are vague and imprecise. We failed to identify an expression closely agreed on by everyone although "low, moderate and high probability" provided fairly good separation. One pathologist stated he always distinguished between "compatible with" and "consistent with" (his ratings 0.4 and 0.8 respectively) when writing reports, but clearly many others make no such distinction and could easily misinterpret his opinion. A radiologist used "consistent with" when certain of his diagnosis, but others showed this could be an expression of low certainty for them. Even physicians rated "pathognomonic" as low as .6 but this may reflect a distrust of others' use of the word.

Numerical scales are not easy to use but clearly they are unambiguous and it is apparent when simple mathematical prinicples are ignored. With increasing interest in formal decision making techniques in medicine, the various scales should give way to a numerical one.

REFERENCES

1. Moore, P.G. and Thomas, H.: Measuring Uncertainty. Omega, Vol. 3, #6, 1975, p. 657-672.

ESTIMATE OF LIKELIHOOD OF PRESENCE OF DISEASE



Physician-Patient Communication:
Interpretations of Non-Technical Phrases

Marcia Z. Wile, Ph.D.,[1] Juanita R. Wingenfeld, B.A., Charlene Ratner, M.S.
Jane E. Mesnard, B.S., Lois P. Mack, M.S., Brian A. Kastner, D.D.S.,
James R. Hickey, M.S., and Steven Dickoff, B.S.

School of Medicine and Graduate School, Case Western Reserve University

Introduction

The importance of physician-patient communication cannot be overemphasized. When communication breaks down, when patients do not understand what is being done for them or what is expected of them adverse consequences may result. A failure to communicate can have detrimental effects on the mental and physical well-being of patients. Therefore, it is essential for health care professionals to become aware of the types of miscommunication that commonly occur (1).

Rapport between physician and patient requires a dialogue which is understood by each individual. Physicians must have an understanding of oral language--how it functions, how it is used, and how it can be misused. Most medical schools teach interviewing skills, but students must also know how language functions in the medical setting. They need to learn to hear what the patient says as well as what the patient means (2).

Patients' misunderstandings of technical language has been well-documented (3-7), but their interpretations of common, non-technical terminology in physician-patient interactions have not been reported. There is a need to determine to what extent problems of communication originate in the non-medical language component of the physician-patient dialogue.

This study was designed to identify the interpretations by both physicians and medical students (referred to as physicians) and potential health care recipients (referred to as patients) of a selected group of phrases commonly used in physician-patient communication. The following questions were asked: Do patients agree on the interpretation of selected non-technical phrases used by physicians? Do physicians agree on the meaning of selected non-technical phrases used in communicating with patients? Is there a difference between patients' interpretations and physicians' meanings of selected non-technical phrases?

Methodology

An instrument was constructed consisting of ten non-technical phrases occurring in physician-patient communication. Multiple-choice alternatives were provided from verbatim responses generated in the pilot testing phase of the instrument. A total of 58 physicians and third and fourth-year medical students (physicians) and 70 potential health care recipients (patients) in Northeastern

[1] Reprint requests : Marcia Z. Wile, Ph.D.
Division of Research in Medical Education
School of Medicine
Case Western Reserve University
2119 Abington Road
Cleveland, Ohio 44106

Ohio completed the instrument. Physicians and medical students were asked to select the response which most accurately indicated the meaning of each phrase, and patients chose the response which represented their interpretation of each phrase.

Percent distributions were compiled for the frequency of physician and patient responses. The Lawshe-Baker nomograph (8) was used to test the significance of the percentages between the two groups. The results indicated considerable variability within each group, and differences between the responses selected by the physicians and patients, some of which reached statistical significance. These results are shown in Table 1.

Results and Discussion

The patients in this study had multiple interpretations of non-technical terminology used by physicians. All but one response for the ten phrases was selected by one or more patients with most agreement being for the phrases, "it's serious," "it's not serious," and "if you have trouble, call me." The overall profile of the patients' response pattern to the remaining phrases showed a range of interpretations.

The results revealed that the physicians participating in the study had various meanings of non-technical language used with patients. Only two of the entire set of responses were not selected. As with the patients, the answers to the phrases, "it's serious" and "it's not serious" showed the greatest congruence among the physicians. The physicians as a group demonstrated more agreement in their responses to the phrases, "if you have a reaction" and "if you have trouble, call me." For these two phrases, most respondents concurred on two of the five selections.

In nine of the ten phrases, some physicians refused to make a selection. They justified their action by stating they either never used such terms in communicating with patients or that the phrases were situation-specific. Most "no responses" were to the phrases, "you don't have to worry" and "take this medicine as needed."

When physician and patient responses were compared, some congruency was found perhaps due more to equally inconsistent response patterns than to agreement between the groups. "No responses" from the physicians accounted for five phrases reaching statistically significance. There were also statistically significant differences in phrases concerning medical management and time.

The question of management and patient compliance arises with the responses to the phrase, "take this medicine as needed." Many of the physicians qualified their responses, stating that medicine would always be prescribed for a specific purpose e.g., pain, nausea, insomnia. However, almost two-thirds of the patients seemed to equate medicine with pain or discomfort. This suggests the need for physicians to be explicit when instructing patients. The potential effects of non-compliance due to misunderstanding or misinterpretation could be serious.

The phrases, "if you have a reaction" and "if you have trouble, call me" had connotations of being situation-specific. Approximately half of the physicians and patients agreed that "if you have a reaction" meant "if something unexpected happens," but the remainder showed little agreement. There was a significant difference in the responses to "if you have trouble, call me." Approximately 60 percent of the physicians, but only 30 percent of the patients interpreted this as "if the treatment doesn't work and you get sicker." It is interesting that 8 percent of the patients felt this phrase was a brush-off.

Those phrases having a temporal component demonstrated the greatest differences in interpretation. The only response to the phrase, "just a minute" that about one-fifth of the physicians and patients agreed upon was "five minutes." Over one-third of the patients interpreted the phrase somewhat cynically as "30 second to 30 minutes," while the physicians were more inclined to answer either a very precise, "a minute or two" or, more realistically, "an indefinite time."

The dictionary definition of "soon" is "at once, immediately, before long, without undue time lapse, promptly, speedily." Yet, almost 90 percent of the physicians interpreted "going home from the hospital soon" as meaning either "2 to 4 days" or "an indeterminate time." However, almost 20 percent of the patients considered this to mean "tomorrow." This suggests that physicians may not say what they actually mean, and patients may not understand what physicians say.

Unlike some of the physicians, all but one patient selected a response for each phrase. Although some had comments regarding the choices or the context of specific phrases, the patients as a group were reluctant to question the validity of the phrases. Despite the fact that there is a period of heightened consumer awareness and increased malpractice suits, the patients in this study seemed very trusting of physicians.

Conclusions

This study has demonstrated the existence of a wide variability in the interpretation of commonly used non-technical phrases within a group of potential health care recipients and a group of physicians, as well as between the two groups. Miscommunication between physicians and patients can have serious consequences for patients, such as undue worry, loss of confidence in physicians, and possible health problems. This study and previous research of interpretations of technical language emphasize the need for explicit communication between physicians and patients.

Educational activities should be initiated at the professional and consumer level to enhance physician-patient communication. For medical students this includes emphasis on precise physician-patient communication and the offering of behavioral sciences courses in medical school curricula. For practicing physicians continuing education programs can be designed to increase awareness of the nature of effective communication. Patient education programs and literature can be created to encourage health care recipients to become more active in their health maintenance and to ask questions when they do not understand physicians' instructions or statements. Only through a conscious effort on the part of physicians and patients can effective communication be realized.

TABLE 1

PERCENT RESPONSE DISTRIBUTION OF PHYSICIAN AND PATIENT GROUPS TO TEN NON-TECHNICAL PHRASES

	Physicians (N=58)	Patients (N=70)	p*
1. Going home from the hospital soon means			
a) in an hour	0.0	3.0	<.05
b) tomorrow	6.0	19.0	<.05
c) after an hour, but today	4.0	7.0	
d) 2-4 days	47.0	45.0	
e) indeterminate time	41.0	26.0	
f) no response	2.0	0.0	
2. Take this medicine as needed means			
a) when you have pain	22.0	36.0	
b) when you feel sick	12.0	16.0	
c) when you feel uncomfortable	31.0	28.0	
d) when you want to	9.0	6.0	
e) as little as possible	14.0	14.0	
f) no response	12.0	0.0	<.001
3. If you have a reaction means			
a) if something unexpected happens	45.0	47.0	
b) fever	2.0	3.0	
c) vomiting	2.0	4.0	
d) itching, breaking out in hives, rash	11.0	26.0	<.05
e) if something happens you don't understand	33.0	19.0	
f) no response	7.0	1.0	<.05
4. It's serious means			
a) if needs immediate attention	90.0	79.0	
b) you'll have to be hospitalized	5.0	16.0	<.05
c) it's cancer	0.4	1.0	
d) you'll need an operation	0.8	4.0	
e) it's terminal	0.8	0.0	
f) no response	3.0	0.0	<.05
5. It's not serious means			
a) the illness is mild	77.0	66.0	
b) it won't kills you	6.0	13.0	
c) it'll go away by itself in a few days	9.0	12.0	
d) nothing is wrong with you	0.5	4.0	
e) no surgery is necessary	0.5	4.0	
f) no response	7.0	1.0	

	Physicians (N=58)	Patients (N=70)	p*
6. You don't have to worry means			
a) everything's fine	17.0	14.0	
b) illness is slight	32.0	29.0	
c) forget about it, learn to live with it	5.0	10.0	
d) false reassurance	4.0	14.0	< .05
e) it's treatable	28.0	33.0	
f) no response	14.0	0.0	< .001
7. I'll only hurt a little means			
a) a quick pinch	14.0	13.0	
b) it'll probably hurt a lot	2.0	20.0	< .001
c) there won't be much pain	34.0	26.0	
d) some discomfort	48.0	41.0	
e) no response	2.0	0.0	
8. Just a minute means			
a) a minute or two	30.0	13.0	< .05
b) five minutes	22.0	21.0	
c) fifteen minutes to two hours	5.0	14.0	< .05
d) anywhere from 30 seconds to 30 minutes	12.0	36.0	< .01
e) an indefinite time	31.0	16.0	< .05
f) no response	0.0	0.0	
9. If you have trouble, call me, means			
a) it shouldn't need further attention	10.0	22.0	< .05
b) reassurance by the doctor	6.0	12.0	
c) a brush-off-doctor doesn't want the patient to call	0.0	8.0	< .01
d) if the treatment doesn't work and you get sicker	59.0	32.0	< .01
e) any reaction - pain or fever	23.0	26.0	
f) no response	2.0	0.0	
10. Take care of it now means			
a) do it the same day	23.0	20.0	
b) within the next few days to a week	12.0	11.0	
c) within two weeks or before the month is over	4.0	5.0	
d) don't put it off	54.0	64.0	
e) no response	7.0	0.0	< .01

*p significance determined by Lawshe-Baker nomograph

REFERENCES

1. Asken M: Communication in a medical setting. J Pract Nurs 28:22-23, 35, 1978

2. Cassell EJ, Skopek L: Language as a tool in medicine: methodology and thoretical framework. J Med Educ 52:197-203, 1977

3. Butt HR: A method for better physician-patient communication. Ann Intern Med 84:478, 1977

4. Fishman SR, Ortiz E: Effective case presentation. Dent Clin 21:539-548, 1977

5. Nichols KA: A reason to worry. Nurs Times 72:1990-1991, 1976

6. Tring FC, Hayes-Allen MD: Understanding and misunderstanding of some medical terms. Brit J Med Educ 7:53-59, 1977

7. Editorial: Hope, but not false hope. Nurs Mirror 142:33, 1976

8. Downie NM, Heath RW: Basic Statistical Methods. Second Edition. New York: Harper & Row Publishers, 1965, pp. 150-151

CLINICAL EXPERIENCE AND THE STRUCTURE OF MEMORY

G.R. Norman, L.L. Jacoby, J.W. Feightner and E.J.M. Campbell
McMaster University, Hamilton Ontario Canada

INTRODUCTION

The process of clinical reasoning, the mental strategies a clinician uses to elicit data from a patient and arrive at a diagnosis is beginning to be understood. The Inquiry project (1), using results from a study of 25 internists, working up actual cases presented by simulated patients, elaborated a four-stage theory whose central element is the generation of multiple competing diagnostic hypotheses early in the encounter. The Clinical Methods study (2) at McMaster replicated these findings, and extended this work to medical students. A significant finding of these latter studies was that the process of clinical reasoning, as described in the Inquiry theory, was essentially similar from the first year of medical school to clinical practice. The one characteristic of the process which consistently differentiated groups with differing experience was the content of the diagnostic hypotheses. The question then remains of how the diagnostic hypotheses are retrieved from memory, and insofar as the hypotheses reflect the structure and content of the clinician's memory, how these structures are formed and modified by educational or clinical experience.

Some insights into these questions can be derived from studies of chess masters conducted by deGroot and others in the 1960's. The thinking process involved in chess playing has some interesting parallels with clinical reasoning. It was found that chess players, like clinicians, generated a number of competing hypotheses about possible moves. The number of hypotheses considered, the depth of the strategy (in terms of number of moves) or the earliness of generating these hypothetical moves did not differentiate chess masters from average chess players. However, in parallel with the medical studies, it was the specific nature of the moves considered which discriminated the master from the average chess player.

Detailed studies of the memory of chess masters revealed an interesting paradox. When permitted to briefly view a typical mid-game position, and then recall the position of individual pieces, chess masters consistently were able to accurately remember the position of more pieces on the board then average chess players, as long as the positions were legal. If the two groups were presented with a mid-game position in which pieces were distributed at random, there was no differnce in the ability to recall the positions of individual pieces.

A parallel situation in medicine is suggested by this finding. It would be expected that an experienced clinician, given a case history of a typical case of some common disease, should be able to recall more details of the case than a neophyte. If presented with a random array of signs and symptoms, however, there should be little or no difference in their ability to recall particular signs and symptoms.

Reprint Requests: G.R. Norman, Ph.D., Department of Family Medicine,
McMaster University, 1200 Main Street West,
Hamilton, Ontario, Canada, L8S 4J9

If such a phenomenon can be identified in the medical environment, there are several potential implications.

1. Detailed study of the clustering and sequence of recalled information may yield insight into the nature of clinical memory.
2. The relationship between measures of recall and the duration and nature of the educational experience may suggest ways in which this critical feature of clinical reasoning can be enhanced.
3. Summary scores derived from the amount of information recalled may provide a simple objective measure of one essential component of clinical reasoning.

With a view to determining whether the finding of deGroot in chess was generalizable to the medical environment, a preliminary study was conducted in the Department of Family Medicine at McMaster. This paper summarizes initial findings from the study.

EXPERIMENTAL METHOD

The basic experimental approach involved administering four written case protocols, two protocols describing relatively typical cases of common diseases, and two protocols which were intended to present an incoherent set of findings not suggestive of any disease process. Each subject was asked to read each protocol throughout at his own speed, as he would a medical record, and then return the protocol to the experimenter when completed. Immediately after reading each protocol he was then asked to recall all the information he read. The total time of reading was noted by the experimenter and the findings recalled by the subject were noted in sequence by the experimenter using a checklist parallel to the case protocol. The subject was then asked to note down his diagnostic hypotheses, and estimate the likelihood of each on a Likert scale. At this point they went on to the next case protocol, until all four protocols had been reviewed.

Order of presentation of protocols was fixed, in the sequence atypical-typical-atypical-typical. The four cases were developed by the physician experimenters; two were based on actual patients, and two were fictitious. The two typical cases were diagnosed as gallbladder disease, and lung cancer. Additional findings were included in each protocol so that every case had a total of 48 available bits of information. In addition, the data were equally distributed across cases in the categories of history, physical, and laboratory investigations.

SUBJECTS

Subjects were chosen at four levels of competence to demonstrate construct validity of the method. Five subjects were chosen from Phase III (year 2) undergraduate medical students, first-year residents in family medicine, third (final) year residents in family medicine, and practicing family physicians with full-time academic affiliation. The mean years in practice of the last group was 15.8 years.

ANALYSIS OF DATA

The analysis of the present paper is focussed on replication of the deGroot studies; i.e. the question of whether experts recall more information for typical cases than atypical cases, and whether this differential recall is related to experience and education. The variables used in the analysis were the number of items recalled for each case, and the time taken in reading the protocol. Because the time of reading the protocol was expected to be directly related to the amount of information recalled, and because there was considerable variation in this time (range = 65 sec → 210 sec), a new variable was formed by dividing the number of items recalled by time spent reading the protocol. Number of items/unit time was calculated for each problem and subject, and the data were analysed using repeated measures analysis-of-variance.

RESULTS

The average number of items/unit time of reading for each problem and group are displayed in Figure I. The expected effect of experience is evident from the figures. No systematic difference between typical and atypical cases is present for medical students or first-year residents. Third-year residents begin to display a higher recall of typical cases, and for the physician group, the effect is strong, amounting to 57% more items recalled per unit time for the typical cases. For the atypical cases, there is no apparent difference in the amount of information recalled by all but the undergraduate students.

To confirm these effects analysis-of-variance was performed, using one grouping factor (educational LEVEL) and two trial factors (TYPE of problem (typical/atypical) and CASE). The results of the analysis are shown in Table I, and confirm the observations of the graph. In particular, there is a significant overall effect of education level - that is to say, increased education and/or experience is associated with a greater ability to recall information. The effect of Type of problem (typical/atypical) is highly significant, indicating that, as expected, subjects recalled more information about typical than atypical cases. The most rigorous test of the research hypothesis is the significant Type x Level interaction, confirming the observation that increased recall of typical cases is associated with increased experience and education. Finally, there is no significant difference between cases, suggesting the effect is not a peculiarity of the cases chosen.

Another way of addressing the question is to determine how many subjects at each level showed the anticipated pattern of recall, with both typical cases leading to greater recall than both atypical cases. No students showed this effect, two first-year residents, three final-year residents, and all physicians. Again, the difference is statistically significant ($\chi^2 = 10.40$, $f = 3$, $p < .05$).

DISCUSSION OF RESULTS

The study we have described has demonstrated a remarkable, yet perhaps puzzling phenomenon. Increasing experience in a domain, be it chess or medicine, leads to enhanced ability to recall the features of typical cases, but no improvement in the ability to recall atypical cases.

From the outset, it is evident that the effect is not simply due to better ability to memorize, since no difference was evident on the atypical cases. Furthermore, although the effect of formal education is confounded with clinical experience in the present study, the strong change over the three post-graduate groups suggests that the instrument is assessing some aspect of memory more strongly related to clinical experience than content knowledge.

Presumably the effect is related to the usefulness of the diagnostic hypothesis as a psychological device for "chunking" information. Yet the relation between the hypothesis and recall is a subtle one. About the same number of hypotheses were advanced by all groups for the atypical and typical cases, although from the structure of the atypical cases the hypothesis would have little utility for economizing memory load. Such is not the case for the typical cases, yet here all subjects (except one medical student) advanced the correct hypothesis, so there was no relationship between advancing the correct hypothesis and success on the test. Thus, although the hypothesis may serve a useful role in clinical reasoning, the ability to recall information was not simply related to retrieval of appropriate hypotheses.

The instrument is clearly in an early stage of development. It has demonstrated impressive construct validity, successfully differentiating groups with presumably comparable medical knowledge but differing clinical experience, an area in which traditional testing methods usually fail. There is a theoretical basis suggesting that the test is assessing factors related to the content of memory. Yet the precise relationship between the measure and experience is unclear. Nor is it evident at this stage how performance on this test is related to accepted domains of clinical competence such as diagnostic reasoning or clinical judgement. Finally, the results are based on a small sample in one, perhaps atypical, clinical specialty, and further application of the method must await replication in other settings with larger samples.

REFERENCES

1. A.S. Elstein, L.S. Shulman, and S.A. Spraflea: Medical Problem-Solving: An Analysis of Clinical Reasoning. Boston, Harvard University Press, 1977.
2. H.S. Barrows, V.R. Neufeld, J.W. Feightner and G.R. Norman: An Analysis of the Clinical Methods of Medical Students and Physicians. Final Report, Ontario Ministry of Health, 1977.
3. A. deGroot: Thought and Choice in Chess. The Hague, Mouton, 1965.

TABLE I

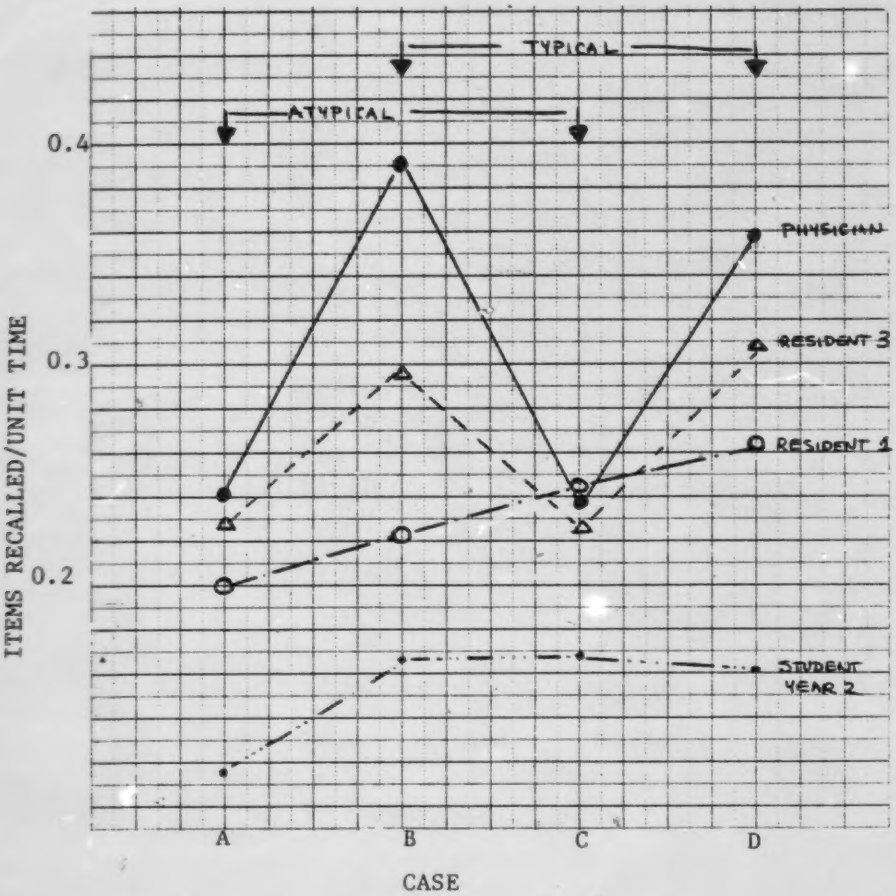
Results of analysis-of-variance with three factors - Educational Level, Type of Problem, and Case.

Source	SS	f	MS	F	p
LEVEL	.257	3	.085	9.28	.001
Error	.147	16	.009		
TYPE	.081	1	.081	40.93	.000
TYPE X LEVEL	.044	3	.015	7.41	.002
Error	.032	16	.002		
CASE	.003	1	.003	1.09	NS
CASE X LEVEL	.010	3	.003	1.06	NS
Error	.051	16	.003		

Other 2-way and 3-way interactions were not significant.

FIGURE I

Items recalled/unit time by physicians, residents and students on typical and atypical cases.



PRECIS

ENHANCING CLINICAL TEACHING

MODERATOR: John Casbergue, Ph.D.
Michigan State University

INFLUENCE OF SEQUENCING IN CLINICAL TUTORIALS FOR SECOND YEAR MEDICAL STUDENTS

Variations in sequences of presentation of clinical tutorials for second-year medical students could not be shown to adversely influence a number of educational outcomes. These outcomes included scholastic achievement, student perceptions of the tutorials as educational experiences, and subsequent choices for clinical training.

DEVELOPMENT, IMPLEMENTATION AND EVALUATION OF A PROGRAM TO IMPROVE CLINICAL TEACHING

This paper describes a program to improve clinical teaching in an Introduction to Clinical Medicine (ICM) course. Program background, program components and evaluation results are reported.

VERBAL PARTICIPATION AND OUTCOMES IN MEDICAL EDUCATION: A STUDY OF THIRD-YEAR CLINICAL DISCUSSION GROUPS

A study of 119 third-year medical students' cognitive level of verbal participation in small discussion case presentation groups was correlated with teacher questions, certain of the students' input characteristics and outcome measures. While there was a correlation between the students' higher level thinking in class discussion with National Board scores and scores on critical thinking tests, multiple regression analyses showed that the students' entering knowledge and skills accounted for the greatest amount of the variance found in the outcome measures leaving classroom processes with little or no contribution.

WRITTEN VS ORAL FEEDBACK: THEIR EFFECT ON LEARNING IN AN INTERNAL MEDICINE CLERKSHIP

Two methods of delivering post-examination instructional feedback were evaluated to assess their effectiveness on subsequent exam performance. Results showed no significant differences in learning could be attributed to either written or oral feedback. Although there was significant improvement on retested items, generalized learning was not enhanced by either subspecialty exams or feedback.

INFLUENCE OF SEQUENCING IN CLINICAL TUTORIALS FOR SECOND YEAR MEDICAL STUDENTS

C.C. Clawson, M.D., Ilene Harris, Ph.D. and Judith Garrard, Ph.D.
Office of Curriculum Affairs
University of Minnesota Medical School, Minneapolis, MN 55455

In most undergraduate medical curricula students are provided with a background in the basic techniques of medical interviewing and physical examination; then they enter a rotating series of clinical experiences which may begin with any of several specialty subjects. Usually the question of which order or sequence of these clinical experiences may be optimal is avoided because factors such as large class size and limited clinical facilities mandate that groups of students must rotate through different sequences. While the importance in curricular design of considering the sequencing of courses and course content is well recognized among education specialists (1-3), little has been done to objectively evaluate the impact of sequencing on important educational outcomes for medical students. Korst and Stone (4) examined the relationship between student achievement in a third-year medicine clerkship and the placement of this learning experience in the year. However, no report available to the authors has examined a broader range of possible outcomes as related to sequencing of clinical specialty experiences.

The present study was undertaken to address two questions regarding sequencing of clinical specialty topics to medical students. First, could variations in the order of presentation of clinical subjects be correlated with any of a variety of educational outcomes. Second, could factors be identified which might serve to minimize any adverse effects of suboptimal sequencing. The focus of this study was a series of clinical tutorials, called Student As Physician (SAP), taken in one of four possible sequences by all second-year students at the University of Minnesota Medical School. The outcomes examined were in three general areas: student achievement, student perceptions of their SAP experiences, and student choices of subsequent clinical training.

THE STUDENT AS PHYSICIAN COURSE AND ITS PLACE IN THE CURRICULUM

In 1969 the undergraduate curriculum of this medical school underwent a major revision (5). Briefly, the revised curriculum consists of: one year of basic sciences and an introductory course in medical interviewing and physical examination, one year of lecture courses based on an organ system format and the concurrent SAP course, and one or two years of elective clinical specialty courses. The SAP course was designed to serve as a transition between the introductory course in clinical data gathering skills and the full-time clinical electives that make up the final two years of the curriculum. The course was conducted in a tutorial format and was included in the schedule throughout the academic year as four half-days a week. The course was comprised of eight specialty topics taught by the faculty of eight clinical departments. The large class size, 240 students, made it necessary to divide the class and provide the topics on a rotating schedule. Fig. 1 shows the four sequences followed by the quarters of the class and relationships of the tutorials to the organ system courses presented concurrently to the full class. The purpose of SAP was to give students ongoing contact with patients and tutor-physicians in a clinical setting through which they could develop and sharpen their clinical skills and apply knowledge gained in the organ system courses. The several specialty tutorials were linked as a single course through a common set of objectives. These objectives emphasized general data-gathering skills and organization, presentation, and synthesis of the data for clinical problem-solving.

METHODS

Subjects. Students entering their second year of medical school were randomly assigned to one of four SAP sequence groups. Nevertheless, the four groups were compared, as described below, for a variety of preentry variables (Table 1). A total of 208 students were included in the study with approximately 50 in each of the four sequence groups. Students who took a reduced course load or a deviant sequence for any reason were not included.

Outcome Variables. Outcome variables were divided into three general categories; 1) student achievement in SAP, concurrent organ system courses, subsequent major clinical courses, and National Board Exams Part I; 2) students' perceptions of the SAP sequences; and 3) students' choices of subsequent clinical training (Table 2). Students' perceptions were obtained by a questionnaire sent to students in their third year, and returned by 87% of the study subjects. Some of the questions included a request for comments or reasons for the primary answers. These comments were systematically categorized by three raters using a coding scheme, which gave greater than 90% agreement when tested for interrater reliability.

Nominal data were examined using contingency tables together with the χ^2 test for statistical significance. Data on an interval scale were tested for statistical significance by analysis of variance. Ordinal data were analyzed by the Kruskal-Wallis test.

RESULTS

Preentry Variables. There were no significant differences among the four sequence groups in any of the following: age and sex, demographic data, premedical college experience, Medical College Admission Test scores, first-year medical school grades, or medical specialty preference (MSP) stated prior to SAP. There were significant differences in work experience. Groups I and III reported teaching backgrounds more often than groups II or IV. Groups I and II reported more work experience in a medical setting than those in groups III or IV. These minor differences among the four SAP sequence groups did not appear to present an obstacle to completion of the intended study.

Student Achievement. Parameters of student achievement analyzed in relationship to the SAP sequence followed were: grades in SAP tutorials, grades for organ system courses, evaluations from third-year clinical electives, and scores from National Board Exams Part I. There were no significant differences among the groups in SAP evaluations. Grades in the organ system courses were based on numerical scores from multiple choice examinations. In the lecture courses groups I and III scored significantly higher ($p < .05$) than students in groups II or IV in the Endocrine and Metabolism course. Group I students scored significantly higher than those in group II in ENT. There were no significant variations in the other organ system courses.

In the clinical electives taken after SAP, a 16-item rating form with a 4-point rating scale was used in evaluation of student performance. Items assessed on the rating form included basic clinical competencies, use of library, attitudinal factors, general medical knowledge, and overall physician potential. Ratings for each of the 16 items were analyzed in the following major clinical electives: Internal Medicine, Neurology, Surgery, Obst./Gynecology, and Pediatrics. The only significant difference was a higher mean rating for case presentations in Surgery earned by students in group I as compared to group IV. Students in the study took the National Board Exams Part I in the spring of their second year. Total scores and subscores were examined in relation to the SAP sequences. There were no significant differences among the groups.

TABLE 1 - Preentry Variables

A. Age and Sex

B. Demographic data
   Hometown size
   Position in family
   Parents' occupations

C. Premedical College Experience
   Colleges attended
   Major field
   Credit hours earned
   Grade point average:
     Science, Nonscience, Total

D. Previous Work Experience
   Medical care setting**
   Teaching**
   Technical
   Dealing with the public
   Other "helping" services
   Unskilled labor

E. Medical College Admissions Test:
   Verbal, Quantitative, Science,
   General information, Total

F. Medical Specialty Pref. (Pre SAP)

G. First-year medical school grades
   Anatomy
   Biochemistry
   Pharmacology
   Embryology
   Histology
   Physiology
   Neuroanatomy
   Pathology

TABLE 2 - Outcome Variables

A. Student Achievement (Grades)
   SAP tutorials
   Organ system courses
   National Board Exams Part I
   Major clinical electives after SAP:
     Neurology, Obst./Gyn., Pediatrics, Surgery**, Internal Medicine

B. Student Perceptions of the SAP Sequences
   Optimal SAP sequence preference; Reasons for sequence preference
   Preparation for subsequent clinical training
     Specific clinical skills**; Specific specialty disciplines**

C. Student Choices for Clinical Training subsequent to SAP
   Clinical emphasis (Track chosen); Reasons for track choice
   Program options: 4-year, 3-year, RPAP**
   Medical Specialty Preference (Changes from preentry preference)

Student Perceptions. The questionnaire was directed at eliciting impressions regarding the optimal sequence for SAP tutorials and the quality of preparation for subsequent clinical electives. In each group about 80% of the students said they preferred the sequence actually followed. Among those who stated a preference for another sequence, there were no significant differences among the groups either in the number of students preferring an alternative or the specific sequence chosen as preferable. The reasons stated for preferring either the sequence actually followed or an alternative choice fell into four main categories: sequence made no difference, 2%; the sequence followed worked well for the student personally without further specification, 3%; the sequence viewed as most advantageous had good correlation with the lecture courses without mentioning any specific lecture course/tutorial matchings, 7%; and responses focusing on one particular specialty tutorial, 88%. Among the latter, 91 students gave reasons related to Internal Medicine and another 27 focused on Neurology as key to their preferences. Pediatrics, Surgery, Obstetrics/Gynecology,

** Items with significant variations were determined between groups (see text).

and Psychiatry were each mentioned by 7 to 12 students as being the course of special concern in their sequence preference. Most students stated advantages to having the particular tutorial of focus either early or late in the year. For example, those who had the major (longer) Medicine tutorial early saw special advantages in this because the general clinical principles and skills that were stressed laid a groundwork for the remaining tutorials. Many students who had the major Medicine tutorial late in the year cited advantages in this arrangement, noting that it allowed them to "put all together" the experience and knowledge gained from earlier tutorials and lecture courses. When tutorials other than Medicine were the focus of students' responses, advantages were noted in having a particular tutorial integrated with a corresponding organ system course. For example, students in group I mentioned the advantage of having the Neurology tutorial in close association with the Nervous System lectures.

On the questionnaire students rated their preparation provided by SAP in 12 basic areas of clinical competence which reflected the overall course objectives of SAP. There were no significant differences among students in the sequence groups in their ratings of preparation in these twelve areas. When the responses from all students were examined the overall mean ratings for each item tended to reflect the degree of emphasis it received in the tutorials.

Students were also asked to rate the preparation for subsequent clinical electives obtained from each SAP specialty tutorial . There were no significant differences among sequence groups in students' judgements of their SAP preparation from the tutorials in Internal Medicine, Surgery, Obst./Gynecology, Family Practice, or Physical Medicine. Students in group III, who had Pediatrics in the first SAP period, felt less well prepared by this experience than students in the other three sequence sections ($p < .05$). Similarly, students in Group IV, who had their Psychiatry tutorial in the first period, felt less well prepared by this experience than those in other groups ($p < .05$). Views of preparation by the Neurology tutorials showed the most consistent pattern. While all students rated their Neurology experience positively, the later in the year that this tutorial was taken, the higher were the ratings given.

Student Choices for Clinical Training Subsequent to SAP. In this school students have 3 options for completion of their undergraduate training after the first 2 years. They may elect either a 3-year program (3 quarters of clinical electives) or the standard 4-year program (5 quarters of clinical electives). Those who chose the 4-year program also had the option of spending their third year in a preceptorship with a practioner in an outlying community, the Rural Physician Associate Program (RPAP). All students chose an area of special clinical emphasis (track) which was entered on completion of the second year and defined the major clinical electives to be taken subsequent to SAP. The tracks offered were: Family Practice, Internal Medicine, Medical Specialties, Pediatrics, Surgery, Psychiatry, Neurological Sciences, and Special or Medical Investigative programs. Students were asked on the questionnaire to indicate their clinical tracks and the reasons for their track selection. There was no relationship between the SAP sequence followed and the clinical track entered. Clinical specialty areas, as reflected in the track chosen after SAP, were compared with the pre-SAP statements of MSP. The track chosen indicated a change in MSP for 23% of the students, but there was no relationship of the changes with SAP sequences.

Reasons for entering specific tracks fell into five specific categories: career goals, 58%; specific positive aspects of the chosen track, 24%; negative aspects of the Family Practice track, particularly the large number of strongly recommended courses, 8%; and a particular faculty advisor, 3%. Only 3 of the reasons mentioned SAP as bearing on the track chosen. The remaining responses were not categorizable. Reasons for choosing specific tracks were not related to SAP sequence. The special programs, RPAP or 3-year option, were examined in relation to SAP sequences. The 3-year option was selected by 22% of the total study group with no relationship to SAP sequence. RPAP was chosen by 23 students. This choice did vary significantly with the students' SAP sequence. Nine students from each of groups I and III chose RPAP, but only two from group II and three from group IV made this choice. Reasons stated for selection of RPAP did not vary significantly with the SAP sequence, but it was noted that groups I and III had much of their primary care experiences early in SAP (See Fig 1).

DISCUSSION

The possible influence of sequencing in a series of clinical tutorials for second year students on several parameters of outcome has been analyzed. The results support, with minor exceptions, the following conclusions:

1) Student achievement as assessed by grades in SAP tutorials, concurrent organ system courses, subsequent clinical electives, and the National Board Exams Part I were not adversely influenced by the sequence followed.
2) There was a high level of student satisfaction with SAP and the preparation provided regardless of the sequence followed.
3) Choice of subsequent clinical training, major curricular options, or changes in MSP all occurred without identifiable influence by a specific SAP sequence.

Exceptions found related mainly to students' perceptions of their preparation in SAP for further training in specific specialties. While the mean ratings were all "good" or higher, and differences between sections were small, there was a tendency to rate higher subjects covered later in the year. This suggested that no matter which tutorial was taken first, the problems of adjustment to the new year altered students' perceptions of their initial tutorial. However, achievement data did not support a predjudicial influence from this.

Among the results, several are worthy of special note. Especially interesting was the high level of student satisfaction with the tutorial sequence actually followed. There was a strong tendency for students to identify advantages in their particular sequence and state these as reasons for preferring that sequence. Objective measurements of academic performance substantiated the concensus of the students' views that there was no "best" sequence among the four.

Those active in curricular design must be cautioned against direct extrapolation of our observations to other circumstances. The findings must be viewed within the context of this particular curriculum. Several factors in the SAP program probably act to minimize potential influence of this example of clinical sequencing. First, while the tutorials are based in specialty departments, all rest on a common set of behavioral objectives that relate to basic clinical problem-solving skills. Second, the organ system courses are taken simultaneously by all students who thus share a common information base. Third, within the SAP sequence there are four tutorials that stress the whole patient and general clinical problem-solving. Thus, while there is rotating scheduling, the student is returning to common and familiar ground repeatedly during the year. Without these features of the SAP course we suspect that the findings of this study might have been quite different.

Figure 1 - Second medical school year curriculum of the study group

Periods	Student As Physician Tutorial Sequences* I	II	III	IV	Organ System Courses
First	Medicine (10 wks) PM&R (2 wks)	Surgery (6 wks) Medicine (6 wks)	Pediatrics (8 wks) Obst./Gyn (4 wks)	Neurology (6 wks) Psychiatry (6 wks)	Cardiovasc. Respiratory Lab Med. Human Sexuality Behavior of Man
Second	Pediatrics Obst/Gyn	Neurology Psychiatry	Surgery Medicine	Medicine PM&R	Gut Man in His Community Blood Skin
Third	Surgery Medicine	Medicine PM&R	Neurology Psychiatry	Pediatrics Obst/Gyn	Fluid/Elec. Kidney Skin Reproduc. Endocrine ENT, Eye
Fourth	Neurology Psychiatry	Pediatrics Obst/Gyn	Medicine PM&R	Surgery Medicine	Nervous System - Muscle Bone -Joint- Connective Tissue

* Family Practice Tutorials were attended for one half-day each month by each student throughout the year.

REFERENCES

1. Posner, G.J. and Strike, K.A.: A categorization scheme for principles of sequencing content. Rev. Educ. Res. 46:665, 1976.
2. Taba, H.: Curriculum Development: Theory and Practice. NY: Harcourt, Brace & World, 1962.
3. Tylar, R.W.: Basic Principles of Curriculum and Instruction. Chicago: Univ. Chicago Press, 1949.
4. Korst, D.R. and Stone, H.L.: The relationship between student achievement and sequencing of student assignment on a Third year medicine rotation. 13th Conf. on Res. in Medical Educ. AAMC; Chicago, 1974.
5. Garrard, J., Lorents, A. and Chilgren, R.: Student allocation of time in a semioptional medical curriculum. J. Med. Educ. 47:460, 1972.

DEVELOPMENT, IMPLEMENTATION AND EVALUATION OF A PROGRAM TO IMPROVE CLINICAL TEACHING

Mark I. Patridge, Ilene B. Harris and Robert A. Petzel
Office of Curriculum Affairs
University of Minnesota Medical School, Minneapolis, Minnesota

Improvement of the clinical education of medical students has received increased attention in recent years. Specifically, efforts have been directed towards the systematic development of clinical courses (Stillman, 1977) and towards development of programs to improve the teaching skills of clinical instructors (Bazuin and Yonke, 1978; Meleca and Schimpfhauser, 1977; Prentice and Metcalf, 1974). This paper describes the development, implementation and evaluation of a program to improve clinical teaching in an Introduction to Clinical Medicine (ICM) course.

Development

Background

It has been demonstrated that medical school faculty have had little or no training in education (Jason, 1977; Page, 1975). Yet, the medical school faculty at the University of Minnesota have expressed great interest in development of their teaching skills. Response to Clinical Teaching problems posed in a newsletter, FORUM, has exceeded all expectations (Patridge, Harris and Masler, 1979); attendance by faculty at workshops on a variety of educational topics has been good; and, participation in a program to improve lecture skills has been enthusiastically received by faculty (Harris, 1979). As part of an on-going faculty development program at the University of Minnesota Medical School, efforts were undertaken to develop a program to improve the teaching of clinical instructors.

First, precepts for clinical teaching were identified by several methods: reviewing the literature of clinical teaching; reviewing students' evaluations of clinical rotations; and discussing strengths and weaknesses of various strategies for clinical teaching with clinical faculty (Patridge, Harris and Masler, 1979). Based on these efforts, a number of factors of effective clinical teaching were identified.

Second, a review of programs to improve clinical teaching was conducted. These programs have taken the form of workshops (Prentice and Metcalf, 1974), videotaping of teaching sessions (Bazuin and Yonke, 1978; Cassie, Collins and Daggett, 1977) and microteaching activities (Littlefield, 1979; Meleca and Schimpfhauser, 1976). Each of these approaches has its advantages and dis-

---

Reprint Requests: Office of Curriculum Affairs
Box 33, Mayo Memorial Building
420 Delaware St. S.E.
Minneapolis, Minnesota 55455

advantages. Workshops serve a number of worthwhile goals such as information dissemination and promotion of attitudinal change (Bland, 1977) but typically do not provide participants with opportunities for practice of clinical teaching skills with feedback. Review of videotaped segments of teaching sessions is an effective approach to improvement of teaching, but this faculty development method is logistically difficult and time consuming with a large faculty. Similarily, microteaching activities are very effective in improvement of teaching. Clinical teaching, however, takes place in a very complex environment, and it is important to observe and assist the clinical instructor in that complex milieu. Yet a recent review of the literature on faculty development programs in clinical teaching, Daggett, Cassie and Collins report that none of the reported programs involve assessment and assistance in clinical teaching based on observations by consultants within actual clinical teaching.

Based on reviews of both the literature on effective clinical teaching and programs to improve clinical teaching, it was decided to pilot-test a program focusing on consultant feedback based on in-person observation of individual instructors in actual clinical teaching sessions. It was also decided to pilot-test this program in the first year Introduction to Clinical Medicine (ICM) course.

There were several reasons for pilot-testing the program in the ICM course. First, during the past five years, ICM tutor commitment to educational development has been demonstrated by their responsiveness to student evaluations of the course and attendance at course orientation sessions. These efforts have been encouraged and given impetus by the course coordinator. Second, since no systematic observation of clinical teaching had been conducted at the University of Minnesota Medical School, it was considered prudent to test the Clinical Teaching Observation Program in a fairly well-structured area of clinical teaching. Although educational consultants in the Medical School were accustomed to observation of lectures as a part of a faculty development effort in that area, few had experience in the observation of clinical teaching for the purposes of the improvement of teaching.

Program Components

The Clinical Teaching Observation Program (CTOP) consists of an orientation workshop, a Pre-Observation Conference between clinical teacher and educational consultant, in person observation of a clinical teaching session by the educational consultant and a Feedback Session. The program is most similar in rationale and design to a clinical supervision model of teacher education (Goldhammer, 1969).

The orientation workshop is held for tutors in the ICM course and addresses topics pertinent to teaching in the course: logistics; strategies for effective clinical teaching; and implications of previous student and course evaluations.

The Pre-Observation Conference is intended to serve as an opportunity for educational consultants to familiarize themselves with the situational characteristics of the tutorial they will be observing and for clinical teachers to identify concerns about their own clinical teaching. The Clinical Teaching Self-Assessment Form (CTSAF) was developed to facilitate productive discussion during the Pre-Observation Conference. The CTSAF consists of a list of 24 clinical teaching dimensions identified through the previously mentioned review of the literature on effective clinical teaching and analysis of students' evaluation comments (Figure 1).

The clinical teacher is asked to rate these teaching dimensions according to their own perceived effectiveness and also their perceived importance. In this way, a focus for observation is identified, since the emphasis is on those dimensions where a discrepancy exists between the perceived importance of a skill and perceived effectiveness in its use.

During the Pre-Observation Conference, a time for observation of the clinical teaching session is arranged. The tutor discusses with the educational consultant which clinical teaching session he/she wants observed and describes the content and setting for that session. The Pre-Observation Conference is a fundamentally important component of the program; feedback is likley to be most useful to faculty if it is based on a contract defining a tutor's concerns.

The instrument developed for use in the observation of clinical teaching is the Clinical Teaching Observation Form (CTOF) which contains a listing of the same dimensions of clinical teaching as the CTSAF. The educational consultant uses the instrument to record impressions about the tutor's effectiveness in each of the clinical teaching skills listed on the observation instrument. Additionally, space is provided for the educational consultant to record evidence of the effectiveness of the clinical teacher's performance.

During the Feedback Session, the educational consultant discusses the observations made during the clinical teaching session using the CTSA form and the CTOF as a basis for discussion. For example, a consultant might make an observation relating to the quality of answering questions:

> *"A student asked you to explain, 'What's a small, medium and large frame?' You pointed to various people in the room and said 'X has a large frame...' It would have been a good idea to make explicit the basis for your characterization or ask the student the reason for her questions so as to respond in terms of her concerns."*

These observations, rich in detail, serve as the focus for discussion.

Before observing tutors and consulting with them, the educational consultants met with the ICM course coordinator to 1) obtain a better understanding of the content and structure of the course, 2) become aware of potential problems they might encounter and 3) recieve a commitment of support from the course coordinator. In order to more fully prepare themselves for the observations, the educational consultants used the CTOF to rate a videotape of a clinical teaching session. Areas of disagreement were identified and discussed; an acceptable level of agreement was reached.

Implementation

During the 1978-79 school year, 30 tutors participated in the ICM course. The course coordinator contacted each tutor by phone, explained the program and asked if he or she would be interested in participating in the program. Emphasis was given to both the voluntary and confidential nature of participation in the program.

Five educational consultants (all non-physicians) and seventeen of the 30 tutors in the ICM course participated in the Clinical Teaching Observation Program. Tutors were randomly assigned to educational consultants; educational consultants would then contact the tutors to schedule the Pre-Observation Conference, the observation and the Feedback Session.

Evaluation

An evaluation form consisting of 22 forced choice Likert type items and six open-ended items was constructed to evaluate the program (Table 1). Of the 17 tutors participating in the Clinical Teaching Observation Program, 13 returned evaluations, resulting in a response rate of 76 percent.

With few exceptions, tutors responded very favorably to the program. Typical of overall impressions of the program were:

> *"An excellent series of sessions. I want to do it again and I would certainly recommend it to anyone else."*
>
> *"I feel that it is worthwhile for us (as teachers) to be looked at objectively by a neutral observer. It makes me appreciate the importance of teaching principles, abilities and strategies."*

The aspects of the program rated highest by tutors were: the emphasis on self-improvement and the individualized observation with feedback from a consultant. Tutors also identified some specific things they intend to do differently with regard to clinical teaching as a result of this program. Specifically, tutors identified feedback as an area for improvement:

> *"...plan better feedback sessions with students."*
> *"...feedback to students, not only immediately but periodically."*

Tutors also indicated an intention to improve their planning for clinical teaching.

> *"...increased consciousness of planning for balanced content."*
> *"...plan time more carefully."*

Negative comments reflected a desire for more critical feedback:

> *"...just generally, a sense of yearning for tougher, more precise critique."*
> *"I asked how I could make myself clear--he said I already was."*

Overall response to the program, however, was extremly favorable. The most frequent comment made by tutors was that one way to improve the program would be to increase the number of observations and consultations.

Discussion

The systematic development, implementation and evaluation of a program to improve clinical teaching has been described. The program consists of a Pre-Observation Conference, direct observation of a clinical teaching session and a Feedback Session.

The program has been well-received by both clinical tutors and educational consultants.

Future plans for the Clinical Teaching Observation Program include the systematic expansion of the program into less structural clinical experiences in the 2nd, 3rd and 4th years of the curriculum. At the same time, the observations and consultations will be repeated with interested faculty in the ICM course.

References

1. Bazuin, C.H. & Yonke, A.M. Improvement of teaching skills in a clinical setting. Journal of Medical Education, 1978, 53, 377-382.

2. Bland, C.J. Guideline for planning faculty development workshops. Journal of Family Practice, 1977, 5, 235-241.

3. Cassie, J.M., Collins, G.F. & Daggett, C.J. The use of videotapes to improve clinical teaching. Journal of Medical Education, 1977, 52, 353-354.

4. Daggett, C.J., Cassie, J.M. & Collins, G.F. Research on clinical teaching. Review of Educational Research, 1979, 49(1), 151-169.

5. Goldhammer, R. Clinical supervision. New York: Holt, Rinehart & Winston, 1969.

6. Harris, I.B. Faculty development program of the Curriculum Affairs Office: Progress report. University of Minnesota Medical School, 1979.

7. Jason, H. Final Report of the Faculty Development Survey. DHEW, Bureau of Manpower, Health Resources Administration, Washington, D.C., 1977.

8. Littlefield, J.H. Improving instructional skills of health science educators. Paper presented at the Annual Meeting of the American Educational Research Association, April 1979.

9. Meleca, C.B. & Schimpfhauser, F.T. A house staff training program to improve the clinical instruction of medical students. In the Fifteenth Annual Conference on Research in Medical Education, Association of American Medical Colleges, November 1976.

10. Page, G., Foley, R. & Pochly, D. A survey of interests in teacher training of health science faculty. British Journal of Medical Education, 1975, 9, 182-187.

11. Patridge, M.I., Harris, I.B. & Masler, D.S. Towards effective clinical teaching: Suggestions for research. Paper presented at the Annual Meeting of the American Educational Research Association, 1979.

12. Prentice, E.D. & Metcalf, W.K. A teaching workshop for medical educators. Journal of Medical Education, 1974, 49, 1031-1034.

13. Stillman, P.L., Ed. Update to Clinical Medicine. Washington: Association of American Medical Colleges, 1977.

14. Stritter, F.T. & Hain, J.H. A workshop in clinical teaching. Journal of Medical Education, 1977, 52, 155-157.

Table 1

Evaluation Item	Mean Rating*
1. This program was valuable to me.	2.15
2. I would recommend this program to a colleague.	2.00
3. It was clear to me that the purpose of the program was for self improvement.	1.69
The following program components were valuable to me:	
4. The Pre-Observation Conference.	2.18
5. Clinical Teaching Self-Assessment Form.	2.92
6. In-person observation of my teaching.	1.92
7. Individualized feedback from consultant.	1.83
The consultant:	
8. Helped me to identify my strengths as a teacher.	2.00
9. Helped me to identify my weaknesses as a teacher.	2.30
10. Provided me with useful suggestions about ways to improve my teaching.	2.00

11. Helped me to explore my feelings about teaching.	2.00
12. Made me feel comfortable.	2.15
13. Was appropriately supportive.	1.84
I feel this program helped me to:	
14. Plan my clinical teaching session more effectively.	2.30
15. Improve my question asking skills.	2.30
16. Improve my question answering skills.	2.46
17. Improve my explaining skills.	2.54
18. Provide better feedback.	2.07
19. Facilitate discussion among students.	2.61
As a result of this program:	
20. I have more confidence in my ability as a clinical teacher.	2.50
21. I have less anxiety about teaching.	2.90
22. I have thought about new approaches to teaching.	2.25

* Mean rating is based on a scale of 1=Very Strongly Agree, 2=Strongly Agree, 3=Agree, 4=Disagree, 5=Strongly Disagree. Response rate=13/17=76%.

Figure 1 : Example of CTSAF Dimensions

University of Minnesota
Medical School

CLINICAL TEACHING SELF-ASSESSMENT FORM

____
Clinical Instructor

Instructions: To improve your clinical teaching, it is first necessary to identify your perceived strengths and weaknesses. This form will serve as the basis for clinical teaching observation and consultation. Please indicate the degree of importance you assign to each of the clinical teaching skills and your perceived effectiveness in each of the clinical teaching skills by circling the appropriate number (when 5 = most important, most effective and 1 = least important, least effective).

IMPORTANCE High Low		EFFECTIVENESS High Low
	ORGANIZATION, PREPARATION, SETTING OF EXPECTATIONS:	
5 4 3 2 1	1. Use of objectives: Exhibits evidence that objectives have been used in the planning of activities.	5 4 3 2 1
5 4 3 2 1	2. Clarity: Provides guidelines concerning goals, student role, level of student responsibility, and the structure of the clinical experience.	5 4 3 2 1
5 4 3 2 1	3. Consistency: Sets expectations that remain constant throughout the clinical experience.	5 4 3 2 1
5 4 3 2 1	4. Level: Sets expectations that are appropriate for the level of the student given the student's academic and clinical background.	5 4 3 2 1
5 4 3 2 1	5. Opportunities to observe: Provides students with opportunities to observe physician-patient interaction and demonstrations by physician.	5 4 3 2 1
5 4 3 2 1	6. Opportunities to practice: Provides students with opportunities to practice clinical skills.	5 4 3 2 1

VERBAL PARTICIPATION AND OUTCOMES IN MEDICAL EDUCATION
A Study of Third-Year Clinical Discussion Groups

by
Patricia J. Foster, Ph.D.
Claremont Graduate School

Introduction

Clinical teaching of medical students takes place in such varied situations as lecture halls, at the patient's bedside and in small group discussions with clinical instructors. Since the ultimate goal of clinical instruction is clinical problem solving (Elstein, Shulman and Sprafka, 1978) considerable effort is (or should be) directed toward improving critical thinking skills. One assumes that the improvement of clinical problem solving skills would most likely take place in the small group where there is close contact between teacher and student. This clinical teaching is usually done by well motivated physicians who have been well trained in their clinical specialty, but have had little formal instruction in teaching. There are few studies of what actually transpires in clinical teaching situations and none of these have related the classroom (or small group) activities to student outcomes. Throughout all levels of education there has been considerable discussion of how "what the teacher does" relates to "how the student does." These "process-outcome" studies are complex and require careful statistical analysis.

The purpose of the present study was to see how much the cognitive level of instructors' questions in small group sessions influenced third-year medical students' responses. Furthermore, did these responses have any measurable relationship to critical thinking skills and National Board of Medical Examiner scores when the influence of certain student input characteristics had been considered?

Review of the Literature

It is intuitively attractive that the more actively the student participates in his learning the more effective it will be. The relatively few well designed studies in this area support this concept (Bloom, 1976; Hecht, 1976; McKeachie and Kulik, 1975; Murray, 1977) but the investigator found no published reports at any level of education that related the amount or quality of the individual student's classroom verbal participation directly to student cognitive outcomes.

The cognitive level of any verbal communication has been classifed by Bloom (1956) according to the amount of integrative function required by the individual to process the information. A simple presentation of facts ("knowledge") is the lowest of the six cognitive levels. "Comprehension" of what these facts mean and their "application" to specific situations requires more complex integration of information. The "higher cognitive levels" of "analysis," "synthesis" and "evaluation" require a more sophisticated integration of the factual information to the particular situation under discussion.

Reprints Available: Patricia J. Foster, Ph.D., Loma Linda University, Loma Linda, CA 92350

The cognitive level of teacher questions and student response both in college and medical school classes is generally at the lowest cognitive level of "knowledge" (Jason, 1962; Barnes, 1975; Foley, Smilansky, and Yonke, 1977). Arnold, Atwood and Rogers (1974) demonstrated that with a minimum of training in the use of Bloom's taxonomy, teachers can raise the level of cognitive functioning in their classrooms. Other studies have shown that students tend to respond at the same cognitive level as the teacher's questions (Davis and Tinsley, 1968; Wood, 1970). Correlational and experimental studies relating the cognitive level of teacher questions to student outcomes show no clear relationship between student achievement and the frequency with which the teacher uses higher cognitive questions (Dunkin and Biddle, 1974; Rosenshine, 1977; Winne, 1977).

Methodology

The sample consisted of 119 third-year medical students and 22 medical faculty in a west coast university. The students were divided by the dean's office into sixteen small groups of six to nine members each to discuss their patient workups. The students' entry characteristics were assessed by three cognitive and three affective input measures. The cognitive measures included an assessment of aptitude--the Medical College Admission Test (MCAT), preclinical performance--The National Board of Medical Examiners, Part I (NBME-1), and critical thinking--The Watson-Glaser Critical Thinking Appraisal Test A (CTA-A). The affective measures were three scales from the California Psychological Inventory (CPI): Achievement via Independence, Achievement via Conformance and Flexibility.

During one quarter of their third year, sixty-two of the small group sessions which averaged 75 minutes each in duration, were observed by the investigator. Each entire session was audiotaped and the teacher and each student clearly identified by notes made by the observer while the session was in progress. Faculty and students indicated that there was either none or minimal observer interference in the group sessions. The observational tool was a modified version of the Flanders System of Interaction Analysis which incorporated the cognitive levels of Bloom's taxonomy in the teacher question and student talk categories. This tool was used to code all the individual student and teacher verbal behavior at three-second intervals throughout each of the sixty-two observed sessions. When all the coding of the tapes was completed, five-minute samples from twelve randomly selected tapes were used for an interrater and intrarater reliability study. The Cohen's kappa for interrater reliability for the twelve sessions was 0.95. The intrarater reliability for the twelve sessions was 0.98. Cohen's kappa was used as the method for measuring agreement between raters because it compares each category unit by unit rather than by total frequencies. With this method observers must agree on almost every event in order to achieve high reliability (Frick and Semmel, 1978).

The outcome measures included the Watson Glaser Critical Thinking Appraisal Test B (CTA-B) and the preliminary ("mock boards") National Board of Medical Examiners, Part II (NBME-2). These data were then tabulated to give descriptive statistics. Product moment correlations and multiple regression analyses were performed to determine relationships between the different variables for the individual students and teachers. The analysis looked at input, process and outcome data separately and in relationship to each other.

One aspect of this study, seldom mentioned in other reports, is the measurement of reliability of the process variables. Reliability was determined by comparing the frequency of the process variables between the first and second halves of the time over which they were studied. Split-half reliability coefficients were obtained by the Spearman Brown formula for each of the process variables. The average reliablity for the major teacher talk variables was 0.80 and for the student talk variables was 0.76. The higher cognitive levels of teacher questions and student discussion were more stable than the lower levels of questions and discussion. "Application" appeared to be the least stable of the cognitive levels.

Findings

Figure 1 shows the distribution of student and teacher verbal participation when averaged over the 62 group sessions. The cognitive level of the teachers' questions and student talk is shown in Table 1.

Figure 1
Student and Teacher Verbal Participation for the
62 Sessions Observed as a Percentage
of Total Class Time

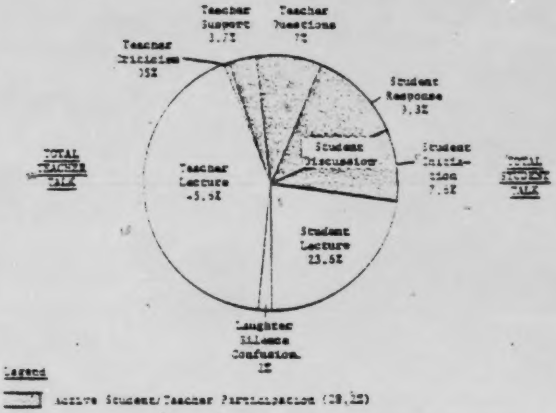


Table 1
Distribution of Teacher Questions and Student Talk by
Cognitive Level for the 62 Sessions Observed (n=119)

Cognitive Level	% of Teacher Question Mean	S.D.	% of Student Discussion Mean	Mean	Product Moment Correlations of Teacher Questions with Student Response (p<.01)
Knowledge	67.2	12.9	75.5	14.0	0.304
Comprehension	21.3	10.5	18.3	11.4	0.309
Application	7.7	7.6	2.7	4.4	0.239
Higher Level Thinking	3.3	3.7	2.3	3.5	0.434
Analysis	2.2	3.1	1.4	3.3	
Synthesis	0.7	1.2	0.3	0.9	
Evaluation	0.4	1.0	0.2	1.2	

Table 1 shows that there was a significant relationship between the cognitive level of the teachers' questions and that of the students' responses. The correlation coefficients for teacher questions and student responses at the cognitive levels of knowledge, comprehension and application were 0.3, while the correlations at the higher cognitive levels of analysis, synthesis and evaluation were more than 0.4.

Correlations between the student's input characteristics and his verbal participation in class showed few significant relationships. Of the 66 correlations that were performed only five were significant. These correlations were low even after they were corrected for the attenuation of both of the variables, suggesting that the underlying constructs did not show strong relationships. Product-moment correlations between the entry characteristics and the student participation variables showed that NBME-1 and CPI Flexibility correlated with the student's initiation of verbal activity ($r=0.27$, $r=0.37$), and higher level thinking ($r=0.33$, $r=0.34$).

The outcome measures, NBME-2 and CTA-B, correlated with the students' higher level thinking talk in class ($r=0.31$, $r=0.41$). When the multiple regression analyses entered the verbal participation variables last into the regression equation it was found that the students' input measures had a much greater influence on outcomes than did their verbal participation in class (Table 2).

Table 2
Summary of Findings for Multiple Regressions of Individual
Student Verbal Participation in Small Groups
with Entry, Outcome and Process Measures
(n=109)

Outcome Measures	Significant Predictors	R^2 Change	F for Variable
NBME Part II Average	NBME Part I Average	.536	97.1
	MCAT General Information	.030	5.1
	Higher Level Thinking	.006	1.4
Critical Thinking Appraisal (CTA Test B)	Critical Thinking (CTA Test A)	.241	17.6
	CPI Achievement via Conformance	.050	4.5
	Higher Level Thinking	.002	0.3

Discussion

The cognitive level of the teachers' questions was significantly correlated with the cognitive level of the individual student's response (Table 1). These findings imply that teacher verbal behavior can significantly influence the student's overt verbal participation in medical school classes. This presumably increases the student's "covert" participation as well, but this was not measured in this study.

The multiple regression analyses show that entry characteristics had a far larger effect on outcome measures than did the process or talk variables in the small discussion groups. There were significant correlations between some aspects of verbal participation and outcome measures (e.g., higher level thinking talk with the CTA-B). The multiple regression analyses, however, "partialed out" practically all of the variance to the entry characteristics leaving verbal participation with no significant contribution to any of the outcome measures. The finding in this study that student verbal participation variables, in themselves, contributed little to the particular outcome measures does not mean that the processes are unimportant, but in the design of the present study they contributed a very small amount to the total variance of these particular outcome measures.

The small discussion group is a part of the medical students' education during their clinical years that has not been previously studied in detail. These case presentation groups are "expensive" in teacher time and while the processes observed did not significantly affect the measured outcomes, these groups provide close student contact with a "role model" who shows them how clinical reasoning is done and how patients' problems can be solved.

While small discussion groups are not unique in higher education, this study shows that in this setting these groups have more student participation, and at higher cognitive levels than has been observed before. The amount and cognitive level of participation is influenced by the teachers' questions, but an even more important determinant of the students' cognitive level of participation is the previous knowledge and abilities they bring to the small discussion group. Since teachers can stimulate more higher level thinking by students, the teacher might be encouraged to use these skills in a conscious, organized manner. This objective as part of an ongoing faculty development program can facilitate more student participation at the higher cognitive levels.

Summary

This study looked at several aspects of clinical instruction in small groups such as teacher questioning and student discussion and documented how certain student cognitive outcomes are related to them.

Drawing from actual transcripts of over 75 hours in small group sessions, the study examined in detail such things as the logic and level of the questions medical teachers ask, the quality of student responses, the cognitive level of classroom discourse and the effects of instruction on selected outcomes such as the students' ability to think critically and their performance on National Board of Medical Examiners examinations. The instrument used for teaching analysis was a modified version of the Flanders System of Interaction Analysis which incorporated Bloom's Taxonomy of the Cognitive Domain.

The cognitive level of the teachers' questions had a significant relationship to the cognitive level of the students' response and discussion. While the students who talked at the highest cognitive level in class were those who scored highest in National Board Examinations (Parts 1 and 2) and the critical thinking tests, the multiple regression analyses showed that classroom activity contributed little to outcomes. The major contributions to outcomes were the aptitude and previous knowledge that the students already possessed.

References

Arnold, Daniel S., Atwood, Ronald K., and Rogers, Virginia M. "Question and Response Levels and Lapse Time Intervals," The Journal of Experimental Education 43:11-15, Fall 1974.

Barnes, Carol P. "A Descriptive Study of the Questioning Behavior of College Instructors." (Doctoral dissertation, Claremont Graduate School, 1975) Dissertation Abstracts International, Vol. 37-01, p. 144A (University Microfilms No. 75-15, 746).

Bloom, Benjamin S. Human Characteristics and School Learning. New York: McGraw-Hill Book Co., 1976.

Bloom, Benjamin S., editor. The Taxonomy of Educational Objectives, Handbook 1: Cognitive Domain. New York: David McKay Co., Inc., 1967. Pp. 144-200; 205-207.

Davis, O., and Tinsley, D. "Cognitive Objectives Revealed by Classroom Questions Asked by Social Studies Student Teachers," Peabody Journal of Education 45:21-26, 1967.

Dunkin, Michael, and Biddle, Bruce J. The Study of Teaching. New York: Holt, Rinehart and Winston, 1974.

Elstein, Arthur S., Shulman, Lee S., and Sprafka, Sarah A. Medical Problem Solving. Cambridge, MA: Harvard University Press, 1978.

Foley, Richard, Smilansky, Jonathan, and Yonke, Annette. "Problem Solvers or Medical Dictionaries: What Are We Teaching For?" Proceedings of the Sixteenth Annual Conference on Research in Medical Education. Washington, DC, November 1977, pp. 233-238.

Frick, Ted, and Semmel, Melvyn I. "Observer Agreement and Reliabilities of Classroom Observational Measures," Review of Educational Research 48:157-184, Winter 1978.

Jason, H. A. "A Study of Medical Teaching Practices," Journal of Medical Education 37:1258-1284, 1962.

Hecht, Lawrence W. "Measuring Student Behavior During Group Instruction," The Journal of Educational Research 71:283-290, 1978.

McKeachie, Wilbert J., and Kulik, James A. "Effective College Teaching," Review of Research in Education, Vol. 3. Itasca, IL: F. E. Peacock Publishers, 1975.

Murray, Michael Dennis. "The Relationship of Classroom Behavior to Academic Achievement and Aptitude." (Doctoral dissertation, The University of Tennessee, 1977) Dissertation Abstracts International, Vol. 12/1A, p. 7156A (University Microfilms No. 780-7711).

Rosenshine, Barak. "Review of Teaching Variables and Student Achievement," in The Appraisal of Teaching: Concepts and Process, ed. by Gary D. Borich. Reading, MA: Addison-Wesley Publishing Co., 1977. Pp. 114-120.

Winne, Phillip H. "Aptitude Treatment Interactions in an Experiment on Teacher Effectiveness," American Educational Research Journal 14:389-410, Fall 1977.

Wood, S. E. "A Multidimensional Model for the Observation, Analysis and Assessment of Classroom Behavior," Journal of Research and Development in Education 4:84-97, 1970.

WRITTEN VS ORAL FEEDBACK: THEIR EFFECT ON LEARNING IN AN INTERNAL MEDICINE CLERKSHIP

John F. Markus, M.S.; Terrill A. Mast, Ph.D.; and Norman G. Soler, M.D.
Office of Curriculum Affairs and Educational Resources
Southern Illinois University School of Medicine

Introduction

Among educational researchers there appears to be general agreement that feedback is a useful instructional technique. So pervasive is this sentiment that many institutions have expended considerable effort implementing feedback into their instructional programs. It is widely believed that by correcting wrong responses made by students, feedback thus enhances their subsequent performance. Kulhavy[1], in his review of "Feedback in Written Instruction," notes that "when students respond to an instructional communication, telling them whether or not their answer is correct increases the amount of material remembered on a later test."

Simply indicating whether or not a student has answered an item correctly does not, however, provide instruction to the student to aid in remediating his deficiencies. Nor does it assist the student who answered the item correctly to transcend his present level of competence. Some instructional information must be included in the feedback if it is to have maximum utility. An unpublished paper from this institution reports that the ideal time for such instructional feedback appears to be immediately after a test, because it is at that time that students are most vividly aware of their deficiencies and are probably more motivated to learn than at any other time.[2]

This research from SIU further suggests using elaborated prescriptive feedback in which students are not only informed whether their answers are correct, but are also given written explanations of why the chosen responses were incorrect and suggested additional readings to aid in remediating deficiencies. In the earlier study, feedback was prescribed to students only for those items they had answered incorrectly. Similarly, Kulhavy points out that feedback is perhaps most useful for those items the student answers incorrectly, but even then is not useful if the student cannot understand the concept and why the correct answer is correct.

Regarding the effects of different modes of delivering instructional feedback most research has dealt with some form of written feedback, e.g., instructional frames in programmed instruction[3-5] or the elaborated feedback used in the project at SIU.[2] A paucity of data exists, however, about how instructional feedback might be effectively delivered orally in the classroom situation.

Interest in this topic was generated in the SIU Department of Internal Medicine when it became apparent that two short subspecialty exams given prior to the final examination were seemingly not as beneficial to students as previously thought. Final exam scores did not reflect much improvement in those subspecialty areas that had been previously tested. While the exams were good

---

Reprint requests should be addressed to: John F. Markus, Office of Curriculum Affairs and Educational Resources, SIU School of Medicine, P.O. Box 3926 Springfield, IL 62708.

indices of student performance at designated intervals in the clerkship, and provided the department with evaluations, their value as instructional tools was questionable. Despite the fact that both oral and written feedback had been provided to students immediately following the subspecialty exams, final exam performance, while adequate, was less than hoped for. Either the students were not availing themselves of the instruction provided by the exams and their feedback, or the instruction being provided was not useful in preparing them for the final exam. An effort was launched to more formally structure the relationship between the subspecialty exams and the final exam, and to maximize the instructional value of feedback after the subspecialty exams.

The purposes of this study were to determine whether giving elaborated feedback on all items to students following an examination facilitated learning, and to determine the relative effectiveness of two modes of feedback delivery--written and oral. In addition, students' attitudes towards both examinations and feedback modes were examined to determine to what extent attitude affected their use of exam feedback and consequently their learning.

Method

Two "subtest" examinations were given during an 11-week internal medicine clerkship involving 20 students. Each exam consisted of approximately 30-35 questions covering three subspecialty areas. Exam One (cardiology, pulmonary, and infectious disease) was given in the third week, and Exam Two (endocrinology, gastroenterology, and rheumatology) was given in the sixth week. Questions for these exams were based on the department's subspecialty objectives and on the subject matter presented in related seminars. At the end of each exam, student answer sheets were collected, at which time the Clerkship Director began an informal review of the first half of the questions. Each question was discussed in some detail. Individual students were called upon to state and defend their answers or explain why they had eliminated other item distractors. The Clerkship Director supplemented the student answers. In this way, correct answers were given, along with feedback as to why each distractor was incorrect.

When half the items had been discussed the Clerkship Director handed out an answer key containing answers to all questions and written explanations for the questions which had not been discussed. The Clerkship Director explained to students that due to time limitations he was unable to discuss all questions and had, therefore, prepared written feedback for the remaining questions. Students were then dismissed. They were allowed to keep their test booklets, answer keys and written feedback, and were not informed to expect retesting on any of those same items.

The final examination, consisting of slightly over 100 multiple-choice questions, was comprehensive, covering an additional three subspecialty areas. Included on the final exam were twenty items from the two previous subtest exams (ten from each). Half of those items were ones for which the feedback had been written; the other half were ones for which feedback had been oral.

On the last day of the clerkship, students were asked to fill out a questionnaire sampling their attitudes about and perceptions of the subtests and feedback. The questionnaire also asked them if, how, and when they had used the subtests and feedback.

Results

To determine whether students performed better on items for which they had practice and feedback than on those for which they had not, a matched-pairs t-test between the scores on the 20 repeated items on the final exam ($\bar{x}=85\%$) and scores on all other items on the final exam ($\bar{x}=71\%$) was computed. This test ($t=5.10$, $p=.0001$) showed that students performed significantly better on the repeated items than on the new items. Further analysis compared student performance on the repeated items ($\bar{x}=85\%$) with their performance on the new items over the same six subspecialties ($\bar{x}=71\%$). This difference was significant ($t=4.27$, $p=.0004$). There was, however, no significant difference found between the performance on new items in the six previously tested subspecialties ($\bar{x}=71\%$) and new items in the three previously untested subspecialties ($\bar{x}=69\%$).

To determine whether written feedback was more or less effective than oral feedback in producing learning, a matched-pairs t-test was run on the subtest-to-final gain on the ten items which had written feedback ($\bar{x}=4.2$) against the ten items which had oral feedback ($\bar{x}=3.9$). This test showed no significant difference in gain due to feedback mode.

To assess the extent to which initial achievement on the two subtests predicted final exam scores, correlations were run between the subtests and the final exam. The correlation between subtest 1 and the final was .50; the correlation between subtest 2 and the final was .24; and the multiple R between both subtests and the final was .56 ($F=3.79$, $p<.05$). The same correlations were run again but omitting those students who scored above 80 percent on either subtest to avoid the ceiling effect. The correlation of subtest 1 with the final dropped to .29; the correlation of subtest 2 with the final was .31; and the multiple R between both subtests and the final dropped to .30, none of which were significant.

Data from the attitude and usage questionnaire was compiled into a composite attitude score for four statements relating to student perceptions of the usefulness to them of both the subtests and their feedback. The mean attitude score was 3.21 (S.D.=.52), on a scale of 1-4, indicating a positive attitude toward the subtests and feedback. A mean score of 2.5 was found on a fifth question which examined student perceptions of the usefulness of the subtests as an evaluation instrument for the department (see Table 1).

The questionnaire also asked students to identify how they had used the subtests and feedback and to identify their preference for either oral or written feedback. All students said that they had used the subtests and feedback for purposes of study or review. Fifty percent said they had used subtests and feedback in reviewing their performance immediately after the subtest exams; 28% said they used the subtests and feedback as study guides prior to the final exam; and 22% said they did both of the above. When asked if they had reviewed all feedback available to them or only feedback for those items answered incorrectly, they all reported using the feedback for every item. When asked for their feedback mode preference, 56% said they preferred written feedback, 16% said they preferred oral feedback, and 28% said they found both feedback modes useful.

To determine if attitude toward the subtests had any effect on learning, a one-way analysis of variance was run on the gain on the 20 common items for the high, medium and low attitude thirds of the students. There were no significant differences between the three groups.

A one-way analysis of variance was also run on the gain on the 20 common items for those who preferred oral, those who preferred written, and those who found both equally desirable. These results were non-significant.

To determine whether student use of the feedback affected learning, a one-way analysis of variance was run between the gain on the 20 common items for three groups: those who had reviewed all items immediately after the subtests to analyze their performance ($\bar{x}=5.8$), those who used the subtests and feedback as study guides for the final exam ($\bar{x}=11.8$), and those who did both ($\bar{x}=7.5$). These results were significant ($F=8.84$, $p=.003$).

Discussion and Conclusions

Several inferences can be drawn from the results reported in this study. First, even though students showed marked improvement when retested over the same items they did not show any evidence of increased general learning in those subspecialty areas. The subtests helped students learn to answer the 20 specific repeated items and therefore we can assume that they also learned to answer the other subtest items which were not repeated on the final. Review of the entire subspecialty area, which we presume occurred both pre-subtest and pre-final, did not have any general effect on learning in that area. This may be due to the breadth of each subspecialty area and the fact that the final exam included only 12-15 questions in each area. In other words, improvement was probably attributable to memory for specific items. Integration of concepts both within and across the subspecialty areas which were tested earlier did not occur.

What we did learn, however, was that the subtests themselves were good predictors of final exam scores. This was a significant finding, implying that the items on the subtests were a good representation of the types of questions contained on the final. This finding also suggests, perhaps, that good students or good test-takers consistently perform better on exams, and that improvement from subtest to final may be more related to individual student ability than it is to content or feedback.

Second, since no significant differences were found in the effectiveness of written and oral feedback and since there was also no apparent relationship between preference for feedback mode and gain on the common items, it is apparent that type of feedback was not a factor in student exam performance. That is not to say, however, that feedback itself was not a factor. Students were positive in their ranking of feedback as an instructional tool and all students said they had used the feedback for purposes of study or review. In fact, our data shows that students studied all feedback available to them, not just items they had incorrectly answered. This is contrary to what we had expected, and contrary to the prevailing belief that students will be motivated only to study feedback for incorrect responses. This finding may be, of course, confined to that particular group of students and may be attributed to their extremely high motivation.

What is perplexing is the significance found in the relationship between feedback review method and gain on the 20 common items. Students who delayed their use of subtests and feedback, reviewing them only in preparation for the final exam, improved their performance more than those students who reviewed the items both immediately after the subtest and in preparation for the final. This finding seems to indicate that improvement was achieved better by pre-final "cramming" than by reviewing feedback after the subtests and "cramming." It may be, however, that a delayed-feedback phenomenon exists and that those students who showed the most improvement benefited from waiting to review the subtest and feedback material.[5]

The effects of this study on the Medicine clerkship are twofold. Because of the generally positive attitude toward the subtests, the department remains committed to administering two subspecialty exams, prior to the final exam, to allow both faculty and students to assess their progress, to encourage distributed study of the content, and to provide a vehicle for discussion of important concepts in those subspecialties. Currently a replication of this study is being conducted to see if the unexpected findings reported reoccur. Continued research on subsequent clerkship rotations will be conducted in an on-going effort to improve the instructional value of the subtests and feedback. One specific issue for future research is the question of generalized learning from the subtests to the final exam.

REFERENCES

1. Kulhavy, Raymond W. "Feedback in Written Instruction," Review of Educational Research, Winter 1977, Vol. 47, No. 1, pp. 211-232.

2. Goetz, Axel A.; Williams, Reed G.; and Boelkins, James. "Elaborated Test Feedback As a Teaching Tool," Southern Illinois University School of Medicine (unpublished manuscript -- available from authors upon request).

3. Anderson, R.C., Kulhavy, R.W., and Andre, T. "Feedback Procedures in Programmed Instruction," Journal of Educational Research, 1971, 62, pp. 148-156.

4. Anderson, R.C., Kulhavy, R.W. and Andre, T. "Conditions Under Which Feedback Facilitates Learning From a Programmed Lesson," Journal of Educational Psychology, 1972, 63, pp. 186-138.

5. Surber, J.R. and Anderson, R.C. "Delay-Retention Effect in Natural Classroom Settings," Journal of Educational Psychology, 1975, 67, pp. 170-173.

TABLE 1

Student Attitudes Toward Subtests and Feedback on a scale of 1 = Strong Disagree to 4 = Strongly Agree.

	Mode	Mean	S.D.
1. Subtests and feedback are useful instructional tools that facilitate learning	3	3.11	0.58
2. Subtests and feedback can be used as study guides for the final exam in Medicine	4	3.56	0.51
3. Subtests give practice on types of items on the Medicine final and other comprehensive exams	3	3.22	0.65
4. Subtests give the student insight into the adequacy of his own data base and the level of his performance	3,4	2.94	0.94
Mean of items 1-4		3.21	0.52
5. Subtests provide the department needed information about students and are therefore useful in discussing mid-clerkship evaluations with students	2	2.50	0.92

PRECIS

ASSESSMENT IN CONTINUING MEDICAL EDUCATION

MODERATOR: Oscar A. Thorup, M.D.
University of Virginia

USING AUDIT AS A NEEDS ASSESSMENT AND EVALUATION FOR A CONTINUING EDUCATION PROGRAM

This paper reports the results of two pilot studies of an audit-based model of continuing medical education. The utility of using audit to identify continuing education needs was demonstrated, but the effect of educational programming on improving performance is not yet clear.

CONTINUING MEDICAL EDUCATION: AN ASSESSMENT OF IMPACT UPON NEONATAL CARE IN COMMUNITY HOSPITALS

Training health professionals in the use of care protocols represented an innovative approach in continuing medical education. The program was designed as a means of enabling physicians and nurses to provide effective, standardized neonatal care in community hospitals. The effects of this educational intervention were assessed by analyzing care process and care outcome.

MEDICAL CARE INTERVENTION: AN EXPERIENCE IN CONTINUING MEDICAL EDUCATION

The educational value of Medical Care Evaluation was studied in thirteen Criteria Development and Audit committees over a seven-month period. The results documented the potential educational effect of the MCE process, and identified several variables affecting their educational value.

Continued. . . .

262

DEVELOPMENT AND EVALUATION OF A COMMUNITY-BASED PULMONARY EDUCATION SYSTEM

A continuing medical education system involving 16 small community hospitals, 8 experimental and 8 control, was developed in which educationally influential physicians were selected from the experimental communities, trained in modern management of chronic obstructive pulmonary disease, and returned to their community. Analysis of their colleagues' behavior from inpatient and outpatient chart audits is being used to determine if this deliberate intervention is an effective new technology for delivering CME to small community hospitals.

PHYSICIAN PRACTICE PROFILES

The practice patterns of Maritime physicians were profiled by compiling listings of every diagnosis managed in the participants' office for a six month period. Results indicate that physician's practices are sufficiently different, but that there is no difference between months, either within an individual's practice or across all practices.

USING AUDIT AS A NEEDS ASSESSMENT AND EVALUATION FOR A CONTINUING EDUCATION PROGRAM

by

Sharon E. Paino, R.N., David W. Cline, M.D., Lawrence Demarest, Ph.D.

Quality of patient medical care has been increasingly emphasized over the past several decades. One method to determine the quality of care is through the use of the medical audit. After constructing criteria about a defined problem area, such as a disease entity, patient records are reviewed to determine deficiencies in patient care. Corrections of some of these deficiencies can be made through education to the health provider. Subsequent re-audit can determine if patient care has improved with further implications about the adequacy of the educational intervention.

The use of the audit as a needs assessment for continuing medical education programs has been sparsley reported in the literature. Eisenberg, et., al.,[1] developed a computer based audit to detect and correct over utilization of laboratory tests. They monitored the use of lactic dehydroginase (LDH) and calcium over a seven day control period, then reported over utilization of the LDH lab test to physicians and provided education of the use of LDH to them. During the control period fifty-one per cent of the charts audited for multiple determinations of LDH revealed over-utilization. During the study period, when physicians were notified if over utilization was found, sixty-five per cent of the charts showed over utilization but this difference was not significant. A simultaneous, undisclosed audit of calcium determiniations also showed no change between the two periods indicating that this method is effective in detecting and measuring over utilization of the laboratory. However, their program of notification and education of physicians was not effective in reducing over utilization.

Escovitz, et. al.,[2] studied the effects of mandatory quality assurance by reviewing hospital medical audit processes. They studied seventeen hospitals in order to determine if those hospitals which had participated in early voluntary efforts to initiate medical audit, differed quantitatively or qualitatively from matched groups of hospitals which had not participated in the previous project. A number of areas were analyzed including the utilization of audit findings to improve patient care. For these variables, no discernible differences were found between hospitals which had participated in the early voluntary project and those which had not. They found that the implications of the audit findings were generally not followed up in an organized and appropriate manner. These findings and others, led Fifer[3] to conclude that the debate persists on whether quality assurance improves care and utilization, cuts costs, changes behavior and is generally useful.

Requests for reprint should be sent to: North Central Regional Medical
Education Center
5445 Minnehaha Avenue South
Minneapolis, Minnesota 55417

These research reports were initial studies that demonstrated early methodological difficulties in the use of audit as needs assessment and evaluation for a continuing medical education program. In order to explore this theoretical orientation more fully, the North Central Regional Medical Education Center (NCRMEC) designed a series of protocols to study this process for the purpose of identifying and resolving problem areas.

Regional Medical Education Centers (RMEC) were established under PL 92-541 and charged to provide Veterans Administration Career Health Professionals with continuing education. The NCRMEC has continuing education responsibility for twenty-four VA Medical Centers in eight North Central states. Rather than approach continuing education from a traditional needs assessment or a "scatter gun" approach, NCRMEC decided to pilot in this unique closed system a continuing education effort based on performance evaluation.

The model used at NCRMEC was of significance because it linked continuing education to actual quality of patient care as suggested by Brown, (1970) in "Mandatory Continuing Education: Sense or Nonsense"[4]. Of particular note is the use of audit to define needs based on performance or behavior rather than knowledge or expressed needs.

The method used by NCRMEC has eight components:

1. choosing a suitable area for performance evaluation from other available data,
2. developing criteria which are predetermined measures of optimal achievable care,
3. data retrieval from the medical record,
4. variation analysis,
5. identification of problems which are educationally remediable,
6. planning a continuing education effort around those problems,
7. conducting the program, and
8. using re-audit six months later to evaluate the effect of the program on the quality of patient care.

Pilot studies using this method were conducted in two different subject areas. The first topic selected for the project was Respiratory Failure. In this first effort, expert clinicians in the field of pulmonary disease were enlisted to develop the criteria and subsequently teach the program. The first pilot involved five hospitals. The criteria set was distributed to the five hospitals and NCRMEC gave assistance with data retrieval. Although fifty records were requested from each facility the total number of records audited in a two year period was one hundred and sixteen. The respiratory failure in these patients was quite severe. Fourty-one per cent (48 patients) died, eighteen per cent had pneumonia, thirty-four per cent had arrythmias, ten per cent had seizures, ten per cent had atelectasis.

The major deficiencies in the audit findings were that:

1. Arterial blood gases were not always done when needed (12%).
2. Central Nervous System depressants were often over used (This varied widely by hospital).
3. The charting was often incomplete, primarily in two areas; indications for intubation and ventilator assistance (4%) and patient education (48%).

4. The rate of correction of hypoxia and hypercapnia varied widely by hospital (These criteria were felt to be unsatisfactory; ignoring many variables and exceptions).

The minor deficiencies in the audit findings were that:

1. Preumothorax was not always promptly re-expanded (4%).
2. Pneumonia was occasionally inadequately cultured.

These variations were analyzed by the faculty and educational consultants to design an education program to address these needs. The program took place in March 1977.

Six months after the course, a re-audit was performed using the same criteria. Four of the five hospitals conducted the re-audit. Hospital five was unable to re-audit because of the lack of patients with respiratory failure in that time period. Several problems surfaced during this project. The major one involved lack of consensus on the criteria from the five medical staffs involved. One facility that showed significant variation from the criteria chose not to participate in the continuing education effort, citing the fact that they didn't concur with the criteria. Of the four hospitals involved in the re-audit, all showed improvement including the hospital that did not participate in the continuing education effort.

Hypertension was the topic for the second pilot project, and the major problems that surfaced in the first pilot were addressed in the second. All twenty-four hospitals in the region were invited to participate in this second effort and each facility was given the opportunity to send a physician considered expert in hypertension to the criteria setting session. Twenty-three of twenty-four hospitals participated in the criteria setting. Criteria consensus continued to emerge as a problem. After three days a core criteria set was developed. Each physician representative was asked to serve as liaison between the NCRMEC and his/her medical center.

The next step was to insure reliability of data by training the committee assistants who abstract the data from the medical record. Each medical center involved sent an audit committee assistant to a special training session.

Several problems emerged at this point. Theoretically, the medical staff agreed to be represented by the physician that attended the criteria setting session. That, in fact, did not occur and, again, physician consensus with the criteria was difficult to obtain at the local level. As a result two major problems emerged: individual hospitals made modifications in the criteria (mostly minor) and only thirteen hospitals elected to continue with the audit. Of these thirteen, four were large teaching hospitals, five were smaller General Medical and Surgical hospitals with minimal medical school affiliation and four were Neuropsychiatric hospitals with medical services.

The thirteen facilities that continued the audit were assisted with data retrieval by a staff member from NCRMEC, and variation analysis was conducted at the local level with NCRMEC participation. The fact that five of the participating hospitals undertook some intervention and re-audited prior to the NCRMEC continuing education effort introduced a confounding variable at this point.

After the data was collected it appeared that even with the modifications and the difference in facilities, patterns of deficiencies in treatment of hypertension emerged. They can be categorized as follows:

1. Inappropriate admissions
2. Diagnosis not always validated.
3. Failure of BP control.
4. Failure to monitor
5. Management of hypertension with renal insufficiency
6. Inappropriate drug combinations
   . Thiazide/Dig. Glycosides.
7. Inappropriate drug doses (low).
8. Obesity ignored.

When the faculty and educational consultants assembled to analyze this data and plan a continuing education program, problems relating to both physician and administrative performance were identified. Therefore hospitals were invited to send a team of both front line care givers and administrative physicians (Chief of Staff or Chief of Medicine) to the program. The program was designed to "train the trainer" and each team was given the task of analyzing their data and then planning an educational intervention at their facility. This local intervention is to be sponsored by NCRMEC. Administrative physician work groups dealt with plans for affecting change at the administrative level (e.g. policy and procedure changes, individual feedback, flow sheets, sanctions).

A re-audit is scheduled for fall, 1979 after the local intervention has occured.

Conclusions

The two pilot tests of this model using medical audit to assess needs for continuing medical education revealed both strengths and weaknesses in this approach. The main strength of the model as a method for needs assessment is that it leads to the design of continuing education which addresses identified deficiencies in the results of medical treatment, rather than basing such education on the opinions of experts, physician perceptions of need, or what other continuing education programs are currently offering. In addition, because standards for evaluation are articulated at the beginning of this process and baseline data are automatically generated, a built-in mechanism for assessing the impact of continuing education programs is provided.

Although the audit-based model has much to recommend it as a needs assessment method, the efficacy of the complete cycle of audit-educational programing re-audit has not yet been demonstrated, as indicated by a serendipitous finding from the first pilot. Since the one hospital which participated in both audits but not in the educational program showed, upon re-audit, progress similar to those hospitals which were involved in the continuing education offering, the effect of the education on the progress made has not been clearly established.

Several explanations for this occurrence are possible. The hospital in question may have taken intervening action on its own. The educational program offered may not have been effective or the participants in the program may not have those who most needed further education in this area. Lack of patient compliance or the action of other care providers (e.g. nurses) may have influenced the status of the criteria. More fundamentally, the deficiencies revealed may not have been amenable to improvement through education, but instead required alterations in the organizations themselves. A final alternative explanation is that conducting an audit and revealing the criteria and findings to physicians may in itself be sufficient to stimulate improvement.

Further research is needed in order to clarify the effect of education in this model. Such research would involve additional sets of criteria (for nursing care and patient compliance) and three groups of hospitals. One group would participate in the total cycle of audit-education-re-audit. The second group would be audited and informed about the findings prior to re-audit, but not take part in the educational program. The third group would participate in both audits, but would not be informed about the results until after completion of the study, nor would they be involved in continuing education.

Even though additional research is needed to clarify the effects of education on the status of audit criteria, the advantages of using medical audits to assess the need for continuing medical education have been shown.

References:

1. Eisenberg, J.M., Williams, S.V., Gardner, L., Viale, R., Smits, H. Cumputer-based Audit to Detect and Correct Over Utilization of Laboratory Tests. Medical Care 15 (11): 915-21, November 1977.

2. Escovitz, G.H., Burkett, G.L., Kuhn, J.C., Zelenznik, C., Gonnella, J.S. The Effects of Mandatory Quality Assurance: A review of hospital medical audit processes. Medical Care 16 (11): 941-9, November 1978.

3. Fifer, W.R. Quality Assurance: Debate persists on goals, impact, and methods of evaluating care. Hospitals 53 (7): 163-7, April 1979.

4. Brown, Clement R., Henry, S.M. Uhl, M.D., Mandatory Continuing Education: Sense or Nonsense. JAMA, Vol. 213, No. 10, September 7, 1970.

Continuing Medical Education: An Assessment of Impact Upon Neonatal Care in Community Hospitals

Wirtschafter, David D., M.D., Sumners, James, M.D., Colvin, Ed, M.D., Knox, Eric, M.D., Turner, Malcolm, Ph.D., Cassady, George, M.D., Beale, Pat, R.N., Budrich, Margaret, R.N., and Brooks, C. Michael, Ed.D., University of Alabama School of Medicine, University of Alabama in Birmingham.

Introduction, Purpose and Problem Statement

Alabama consistently ranks among the states with the highest neonatal mortality rates (1976: Ala. = 12.3/1000; U.S. = 10.7/1000). From the UAB perspective as one of the state-wide referral centers for sick newborns, it is evident that the quality of neonatal care suffers not so much from a lack of physician attendance at deliveries but from a lack of prompt and appropriate postnatal medical process. The purpose of this paper is to describe the development, implementation and evaluation of a method of continuing medical education designed to increase the quality of neonatal care delivered by practitioners in the community. It presents an assessment of an innovative approach to the dissemination of medical information to physicians and nurses in community hospitals and its impact on the most visible of patient outcomes: mortality.

Specifically, a two-fold problem was addressed: 1) a study of the continuing medical education and care protocol method for improving the neonatal care delivered by community-based physicians (general practitioners and pediatricians); and, 2) an assessment of the impact of the care protocols to increase the effectiveness of neonatal care in the community hospital setting.

Background and Review

Provision of neonatal care represents a unique setting for the design and development of alternative health care delivery strategies. The neonate's outcome is unaffected by the patient compliance factor but is, however, variably affected by the access factor. An important consideration for neonates is whether there is access to fully qualified medical process or merely access to individuals certified to deliver medical care process. Data in Alabama indicated major deficiencies in the capability of community hospitals to deliver optimal neonatal care. The development and use of neonatal care protocols (e.g., clinical algorithms) represents an attempt to institute a consultant-extender system that would enable the broadly trained practicing physician to recognize and stabilize neonatal diseases at the specialist's level.

Address for reprint requests:
David D. Wirtschafter, M.D.
Clinical Information Systems
809 - 19th St., South
University Station
Birmingham, AL 35294

This project was supported by a Bureau of Community Health Services (DHEW) Grant (MC-R-010381)

Care of the sick neonate has undergone vast changes in the last two decades. Cassady (1) has reviewed the key innovations and their effects on neonatal mortality rates. As impressive as these changes are, it is apparent that the cost of delivering such services within every hospital unit is prohibitive, and that medically trained manpower is simply unavailable. The practice of prompt referral to a neighboring neonatal intensive care unit has accordingly become commonplace.

Usher (2) has analyzed the frequency and preventability of the perinate's problems. For instance, his analysis of neonatal mortality in Quebec led him to assert that there is an "irreducible" minimal neonatal mortality of 2.6/1000 live births greater than 1000 grams, composed almost entirely of anomalies and Rh deaths prior to week 32. He attributes regional and hospital differences in perinatal mortality rates to variations primarily in the neonatal mortality among low birth weight infants, that is, to the inability of hospitals not equipped with neonatal intensive care units (NICU) to cope with selected neonatal medical problems.

A survey of labor and delivery room professionals in Alabama hospitals delivering more than 500 babies per year indicated that about half of the infants were delivered by obstetricians and the remainder were delivered by general practitioners. One-quarter of them reported that the mother's prenatal records were not available during labor. One-half of the respondents reported that both one and five minute Apgar scores were routinely judged; all but one utilized a wall clock rather than a stop watch/timer to ascertain the time. A majority of hospitals had two nurses routinely to dry the infant within one minute of delivery. Respiratory support equipment was generally complete. The most outstanding deficiencies were apparent in the lists of resuscitation medications, but data indicated that active resuscitation was carried out in infants weighing less than 1.5 kg; however one in every three responding hospitals indicated that this was not the case for infants weighing less than 1.0 kg. The available data about preparations for neonatal care among our referring hospitals, although anecdotal, suggested major deficiencies in the community hospitals' capabilities to deliver optimal neonatal care.

Neonatologists have thus far been focusing on improving the capability of community hospital facilities and personnel. This approach, a traditional one, has been one of post-graduate medical education; the consultant-extender concept, which involves the use of neonatal care protocols, represent a complimentary approach to the more familiar modes of continuing medical education.

Objectives

The consultant-extender system required that the clinical decision-making process be specified in an objective manner. The first objective of the project was the development and testing of clinical algorithms for seven (7) selected neonatal problems.

A second objective was to introduce the consultant-extender system and train participants in such a manner that the impact of the project could be evaluated. The implementation strategy was designed to achieve randomization of the many variables affecting the outcome of perinatal care.

The third objective was to evaluate the consultant-extender system in terms of its four facets: 1) its validity as an instrument to convey optimal neonatal medical decisions; 2) its suitability as an operable instrument in managing patients; 3) its effects on the processes of neonatal care; and 4) its effects on the outcome of care.

Development and Design

Neonatal care protocol development involved the following activities: 1) selection of "management problems"; 2) description of the care process for each problem; 3) conversion of the care process to logic diagrams; 4) design of care protocol documents that incorporated only those information elements necessary to communicate the medical logic for managing the selected problem; 5) testing of the care protocols for medical accuracy and ease of utilization; and 6) printing and distribution to all selected users.

Protocols were developed for the following neonatal problems: 1) respiratory distress; 2) jaundice; 3) asphyxia; 4) LBW stabilization and follow-up; 5) neonatal seizures; 6) apnea; and 7) macrosomia. Validation of the neonatal care protocols encompassed both retrospective and prospective phases.

Implementation Activities

The implementation activities of the program involved the following tasks: 1) preparation and approval of human use/informed consent agreements; 2) state-wide recruitment of hospital staffs; 3) allocation of participating hospitals to study groups based on statistical analyses; 4) development of protocol training methods and materials; and 5) development of a database containing selected birth, medical care process, and outcome variables.

Hospitals that elected to participate in the study were allocated to one of three groups on the basis of the following variables: 1) total births, 2) low birthweight births (LBW); 3) low birthweight deaths; 4) black births; and 5) referral rates.

Group 1: Experimental group (N = 13)- This group received an incremental education program (mediated lecture/case studies) combined with installation of neonatal care protocols into each hospital.

Group 2: Placebo group (N = 15)- This group received an incremental education program (mediated lecture/case studies) but no installation of neonatal care protocols into the hospitals.

Group 3: Control group (N = 13)- This group had no incremental education program nor were care protocols installed into the hospital.

Results

The care protocol system was expected to result in two effects: 1) an improvement in the quality of the neonatal care in the community; and 2) an increase in the referral rates from hospitals who were not actively participating in the regionalization program. Acceptance of the protocols was measured by two parameters: first, the number of cases utilized to any degree was compared with the number of applicable cases; and, secondly, a chart audit procedure indicated the number of times recommendations were followed as compared with the number of recommendations made. Data evaluation was directed to the following four hypotheses.

Hypothesis 1: There is no significant difference in the frequency or severity of risk factors among the three study groups.

A comparison was made for the following risk factors: birthweight, gestational age, and relative mortality risk (RMR) scores. An analysis (t test) indicated that there were no significant differences in birthweights among the three hospital groups. Gestational age was significantly greater in the protocol group than in the control group but less than that found between the lecture and control group. The relative mortality risk (RMR) was based upon neonatal mortality experience in terms of the birthweight and gestational age of all admissions. Analysis (t test) of the RMR scores indicated that the protocol group had a significantly higher risk population than either the lecture or control groups, however the lecture group was significantly more at risk than the control group. The protocol group had a significantly higher incidence of infants with predictably higher risks of neonatal mortality.

Analysis of the obstetrical risk factors among the three groups was determined by the Hobel Risk Score Method. Analysis (t test) of the mean antenatal and intrapartum risk scores showed no significant differences between the three study groups. These data indicate that the protocol group did not contain mothers with significantly fewer risk factors than the lecture or control groups.

Hypothesis 2: There is no significant difference in the adequacy of antenatal and intrapartum care among the three study groups.

The results from chart audits provided data on the frequency of fetal heart rate monitoring. Chi-square analysis indicated that fetal heart rate was monitored significantly more often by the lecture group; monitoring fetal heart rate by the protocol group was significantly greater than the control group but significantly less than that reported by the lecture group.

Hypothesis 3: There is no significant difference in the quality of neonatal care among the three study groups.

An attempt was made to ascertain differences between the three study groups based upon neonatal care process. Chi-square analysis indicated no significant differences in the incidence of respiratory distress syndrome (RDS). An analysis was conducted on all infants who exhibited symptoms or signs of RDS without any consideration to treatment either by referral to a NICU or treatment in the community hospital. Sub-groups using the RDS protocols produced significantly better documentation of the clinical conditions based upon information on the patients charts. Analyses also were conducted related to other variables such as documentation of temperature recording, initial temperature, blood pressure, glucose levels, hypoglycemia therapy, blood ph, incidences of acidosis, chest X-rays, and referrals to NICU's.

Hypothesis 4: There is no significant difference in low birthweight (LBW) neonatal mortality rates among the three study groups.

No significant differences were found between the groups based upon the total number of births, the total number of low birthweight births, the total number of low birthweight deaths, the number of black births, and the number of referrals for each of the hospital groupings. A comparison of the crude LBW neonatal mortality rate among the three hospital groups indicate that dramatic reductions have occurred in neonatal and low birthweight mortality rates during the last three years. However, it is also equally evident that belonging to the "protocol" group or the "lecture" group did not have a distinguishable effect over and above belonging to the control group. All groups met or exceeded the goal of reducing their low birthweight neonatal mortality rates by at least 25%.

Regression Analysis

Extensive regression analyses of the total study group were performed. Variables were analyzed by an "all possible regression subroutine" and from these, the final variables were selected. There were no significant differences found between the control group and the protocol group. The lecture group tended to have a higher mortality rate and the relationship between belonging to the lecture group and poor neonatal outcome is not readily explained. The chart audit group analyses showed essentially similar results; there were no significant differences between the three groups. A separate subset (audit group) was defined by the presence of respiratory distress symptoms, signs, or oxygen therapy. Sequential analyses were performed and the results were consistent with other findings in that there were no significant differences. The relative post-delivery mortality risk factor was the only variable consistently related to neonatal survival. Membership in any one of three hospital groups was not significantly related to neonatal survival.

Discussion

The continuing medical education program consisted of an integrated didactic, slide-tape presentation on the recognition and stabilization of selected neonatal problems. This presentation and subsequent information provided in the Alabama Perinatal Bulletin was an attempt to disseminate current knowledge about the recognition and stabilization of infants with neonatal problems to as many physicians as possible.

Care protocols were successfully developed for seven problems associated with neonatal morbidity, however, hospital acceptance of the protocols was modest. In one-half of the protocol group they were never used and in the other half, protocol usage varied from 9% to 62% in the potentially applicable cases. This usage pattern reflects a lack of positive feedback and reinforcement in the project design.

Attention was directed to the question of whether protocol hospitals demonstrated significantly different processes of neonatal care even though the "outcomes" were not different. A comparison between the protocol group and the lecture and control groups indicated that host factors, as judged by the initial severity of disease, were substantially similar in all three groups. However, the evidence of medical intervention in these three groups showed significant differences. In the protocol group, blood pressures, glucose levels, and blood gases, were measured more often and chest X-rays were also done significantly more often. Protocol hospitals produced closer adherence to accepted standards of care process. The fact that medical care evaluators have had and continue to have great difficulty in demonstrating statistically significant effects of process adherence with favorable outcomes is an enigma of the field, and it does not diminish the potential significance of obtaining process adherence to recognized standards for managing disease. The fundamental problems of evaluation continue to elude solution. Neither the literature nor our own dataset has enabled the development of a satisfactory set of variables to prognosticate neonatal outcome on the basis of clinical parameters or educational intervention. If such a prognostication score could be successfully identified and validated, then the capability of assessing the impact of clinical and education interventions on care process and care outcomes would be greatly enhanced.

The project goal was to achieve a 25% decrease in the low birthweight neonatal mortality rate and evaluate the efficiency of this intervention as differentially impacting on the rate of decrease in the neonatal mortality rate. Decreases ranging from 25% to 40% occurred within the study groups. These represent annual changes of 7% to 13% as compared to the previous decade's average of a 3% annual decrease. It is evident both by simple statistical analyses and by the multiple regression analyses that the continuing medical educational components of the program had no differential impact in affecting the resulting changes in neonatal mortality rates. This is a noteworthy finding in the evaluation of continuing medical education programs, as this program was more intensive and directed than others.

REFERENCES

1. Cassady, G.: The Impact of Perinatal Intesnive Care, in Clinical Perinatology. A. Brown and S. Aladjem, eds. Mosby Company, In Press.

2. Usher, R.H.: Clinical Implications of Perinatal Mortality Statistics, Clin. Obstet. Gynec. 14:885, 1971.

MEDICAL CARE EVALUATION: AN EXPERIENCE IN CONTINUING MEDICAL EDUCATION*

L.J. Sandlow, M.D., P.G. Bashook, Ed.D., J. Maxwell, M.A.
Educational Development Unit
Michael Reese Hospital and Medical Center
Chicago, Illinois

The educational potential of Medical Care Evaluation (MCE) has been in two areas: as a means of identifying physicians' educational needs and as a continuing medical education (CME) activity for physicians.

The potential of MCE as educational needs assessment has been recognized for some time. In his well-known Bi-Cycle model, Clement Brown, Jr., conceptualized the needs assessment function of MCE as educational diagnosis.[1] Through MCE, physicians' educational needs can be diagnosed on a local basis, and CME programs, tailored to meet these needs, are the educational therapy. The Bi-Cycle model has been widely endorsed. It is the focus for the professional Standards Review Organization's MCE procedures.[2] The model is also used by the Joint Commission on Accreditation of Hospitals in its Professional Evaluation Procedure (P.E.P.) manual and workshops.[3] Before discussing these educational potentials further we will briefly describe the MCE process.

THE MCE PROCESS

The accepted goal of all MCE programs is the maintenance and improvement of the quality of patient care, but each program emphasizes the components and issues most appropriate to local goals and setting. The following figure and text briefly describe the MCE process from the perspective of the Michael Reese Hospital program[4] (see Figure 1).

Problem Selection is accomplished by group assessment of (a) the current morbidity and mortality associated with the problem, (b) its impact on the health care system, (c) the potential for improvement considering the current state of the art, and (d) the use of the problem as a tracer of health care in the setting being examined. Because MCE studies are in-depth and require much physician time, the paramount concern is identifying problems for which the MCE process will yield the greatest effect on quality of care.

Criteria of Care are developed for each of the health problems selected; again, by the use of a group process. Criteria are formulated for both processes and outcomes of health care. The role of the criteria is to separate the simple, appropriately treated cases from the complex and the inappropriately treated cases, rather than to constitute minimal standards that could be interpreted as "cookbook medicine."

---

*The research reported in this paper is being supported by a grant from the National Fund for Medical Education.

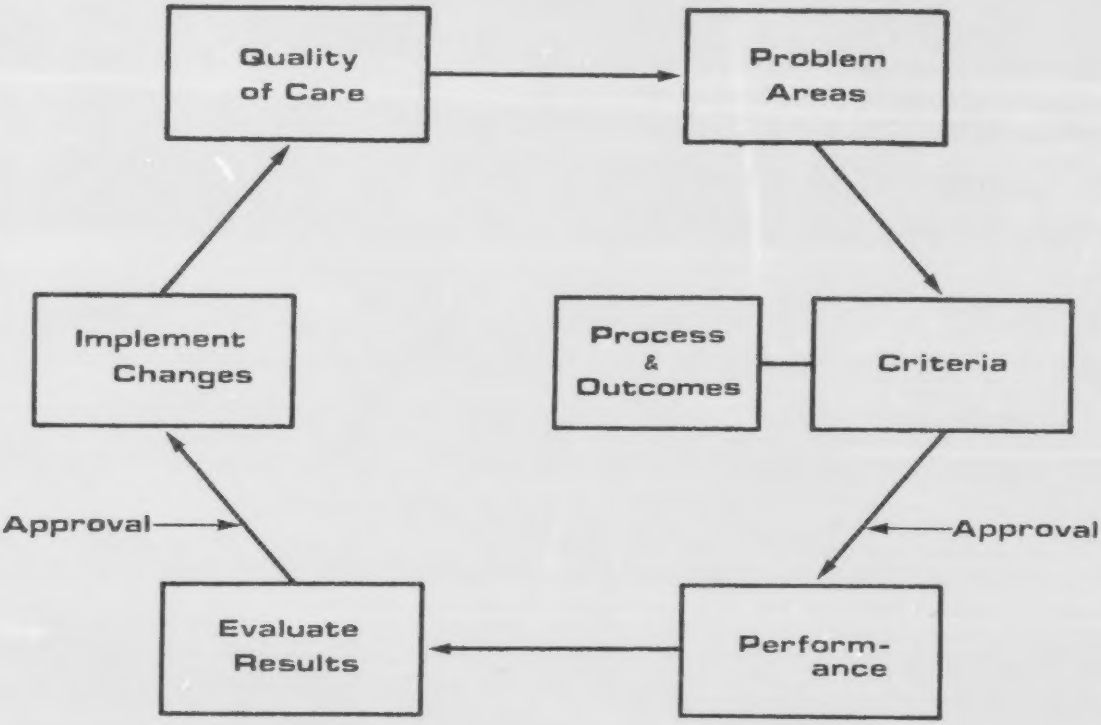
Send reprint requests to Leslie J. Sandlow, M.D., Educational Development Unit, Michael Reese Hospital and Medical Center, 530 E. 31st Street, Chicago, Illinois 60616

Measurement of Performance involves defining criteria in explicit terms to enable non-physician abstractors to determine from the patient's record whether each criterion has been met or not, without having to make clinical judgments. The performance data are summarized in a format which allows a physician group to review compliance for each criterion as well as for each individual case.

Evaluation of Care is performed by a physician group using the performance data for the health problem being studied. First, the group evaluates overall care by comparing actual performance against the criteria; second, the group examines individual records selected because of lack of compliance with key criteria. Because the criteria were developed to separate simple appropriately treated cases from complex or inappropriately treated cases, the physicians review the entire course of hospitalization or outpatient care for selected cases, with the emphasis on determining appropriate management and problem solving, rather than compliance with the criteria. Third, the group discusses recommendations to improve the care and the audit process.

Implementation of Recommendations are carried out, usually by the clinical departments or institutional administration. Recommendations fall into three areas: professional education (the needs assessment function described previously); new patient care and patient education programs; and administrative and structural changes.

Figure 1. The MCE Process



THE RESEARCH QUESTION

There is evidence that MCE is not successfully identifying physicians' educational needs. Sanazaro reviews several studies which applied the Bi-Cycle model.[5] Deficiencies in physician performance were identified, CME programs implemented, and subsequent physician performance evaluated through record audit. The results were not impressive. In some hospitals, there was no improvement in physician performance following the CME intervention. In others, the improvement was uneven. The largest improvement reported was at the hospital where the Bi-Cycle model was first implemented. In each of three health problem areas, there was only a 50 percent improvement in physician performance after two years of successive CME and medical record audits. At Michael Reese Medical Center, over 185 MCE studies in more than 120 health problem areas have been performed. Not a single one of these studies has identified deficiencies in physician performance which would be remedied by a formal CME program.

It appears that the educational potential of MCE as needs assessment is not being realized. On the other hand, there is a growing awareness that the MCE process itself serves as an effective educational experience and a stimulus to improve performance for the physicians who take part in it. In fact, this view was held by the early medical record and peer review committees who justified their activities as CME.[6]

Ashbaugh and McKean also support this view on the basis of their experiences with MCE.[7] The initial goals of their MCE program were based on the Bi-Cycle model of educational diagnosis and therapy. Deficiencies were identified and educational programs initiated. They found, however, that the audit process itself was the "most educational experience of all." Because their MCE committees felt that standard criteria were too cumbersome, they developed their own and discovered a beneficial side effect. The process of criteria development was itself one of the key educational experiences in the audit.

In the Sanazaro review, four studies are described which provide further experience that MCE is effective CME. In each study, physicians involved in some aspect of MCE exhibited substantial improvements in performance. These improvements were specific to those patient care topics on which the physician had focused during his/her participation in the MCE process. As yet, however, the educational effectiveness has not been demonstrated experimentally.

The potential of MCE as CME has been observed informally and numerous anecdotal reports have been recorded. As yet, however, no one has either documented the educational value of the process or attempted to design a systematic CME program within the MCE process. We have implemented a study to accomplish both aims. This report is devoted to documenting the educational value of the previously described MCE process. A future report will be devoted to measuring the educational value of a formally designed CME program.

METHODS

A trained observer witnessed 13 different MCE committees at Michael Reese Hospital over a seven-month period, involving 25 one-hour sessions of seven criteria development committees and 25 sessions of six audit committees. Attendance averaged between three and five attending physicians (usually from the same specialty), with one serving as committee chairman, a MCE staff leader (always a physician), a medical record specialist and observer.

The average physician membership of most committees was between four and eight, excluding MCE staff; some committees included as members one or two post-doctoral fellows or residents, nurses, and allied health professionals.

The only exceptions were two committees, Emergency Medicine and Pediatric Acute Care, where committee composition was primarily nurses, residents and allied health professionals.

Of the 13 committees, nine derived from seven clinical departments, were observed in three or more sessions. They comprise four criteria development committees (two with inpatient care emphasis, and two with outpatient care emphasis) and five audit committees (all with inpatient care emphasis). Findings are based mainly upon observations in these nine committees.

Data collection methods used to document naturally occuring CME in these committees were:

1. Systematic Observation of Committee Sessions

Using a coding scheme, a trained observer attempted to classify events in the meetings which could be considered potential "learning opportunities." The coding categories focused attention upon how and by whom a potential learning opportunity was initiated; whether the content was primarily problem definition, diagnosis, therapy, actual patient care practices, etc; whether presentation was by a single speaker or as a discussion; and its duration. Data were also collected on characteristics of committee chairmen, the MCE staff leader, committee members, and committee functioning which might affect the educational value of meetings.

2. Semi-structured interviews with eight physician members of committees

The interviews averaged 30 minutes each and explored the educational value of MCE, the structure and operation of the committees, and the physicians' attitudes toward MCE. The data served to check the validity of observed "potential learning opportunities" and will be used in the second report of the study for assessing changes in physician attitudes.

3. Informal discussions with committee members and MCE staff

Anecdotal information about committee functioning and the educational value as depicted by informants served to refine, focus, and enhance the data obtained from observing committee meetings.

The data collected by these three methods was systematically organized and summarized to derive characteristics of a committee functioning with potentially high or low educational value.

RESULTS

Almost all of the committees could be clearly defined as having either relatively high or relatively low educational value. Of the nine committees studied in depth, three were judged to be of high educational value, while six were judged low in educational value. The single characteristic which most clearly separated committees with high educational value from those with low educational value was the number and duration of discussions per meeting that concerned information on the problem and its appropriate management. In committees of high educational value, the average number of such discussions per 1-hour meeting ranged from two to five for the different committees, and the average total time per meeting spent in learning opportunities ranged from 6 to 18 minutes. In contrast, in the committees judged to have low educational value, the average number of such discussions per meeting ranged from 0 to 0.7 for the different committees, and the average time per meeting spent in learning opportunities ranged from 0.8 to 4.2 minutes.

The committee chairman has a major effect on the educational value of the committee meetings, more through his or her attitude and goals than through the way the meeting is actually run. Of the three committees with high educational value, two have chairmen who clearly see the committee as having an educational function, while none of the committees with low educational value have chairmen with this attitude. Of the former committees, one is to a significant extent "taught" by the chairman, and the second has been structured by the chairman to be educational without requiring major direction from him; the third is educational because of the attitudes and approach of the members, rather than the direction of the chairman, although his views may have been important in creating these attitudes.

There appears to be a minimum "critical mass" of attending physicians required to sustain an educationally valuable interaction. This "critical mass" varies with the characteristics of the committee members, but all six committees which had a regular attendance of less than four attending physicians were judged to be of low educational value, while only one of the seven committees with an average attendance of four or more attending physicians was of low educational value.

In terms of their methods of operation, committees with high or low educational value can be characterized as follows:

Criteria Development committees with high educational value focus on the optimal, as well as the minimal acceptable, management of the problem under consideration. There is an exploration of the nature of the problem and alternative approaches to diagnosis and treatment, and a spontaneous presentation of opinions and the justification for these. Members often support their views with references to the research literature.

Criteria Development committees with low educational value focus on the minimal, agreed-upon standards of acceptable care, and most of the points raised in meetings are ones for which there is already agreement.

In Audit committees with high educational value, records are reviewed by one member prior to the meeting, and are presented as case problems. These are then discussed by the committee, with the primary concern being the accuracy of the diagnosis and the overall appropriateness of the care. The criteria which the record failed to meet are seldom explicitly mentioned, unless the record raises a question about their validity.

In Audit committees with low educational value, records are reviewed (in some committees, during the meeting itself) primarily to see if the criteria were actually met or not. Cases are rarely discussed unless a member asks for an opinion as to whether or not something is appropriate or justified.

DISCUSSION AND IMPLICATIONS

The above results demonstrate both that MCE committees have potential educational value, and that differences in the structure and functioning of the committees have an important effect on this value. Considered as CME, MCE committees can be described as follows:

1. Criteria Development Committees are health-problem-oriented conferences which explore the nature, diagnosis, and treatment of the problem under consideration.

2. Audit Committees are patient-management-oriented conferences which review the overall appropriateness of the care received by the patient whose record is being reviewed.

Based upon principles of adult education,[8] the potential for CME in the MCE process seems apparent. Participating physicians are not passive recipients of information. Rather, they are actively involved in problem identification, definition and resolution. The committee members have the opportunity to thoroughly explore the current state of medical knowledge with respect to a specific health problem. Furthermore, the MCE process naturally assures that learning takes place in a meaningful context: the question of what constitutes appropriate care of a patient.

The MCE process, to become beneficial CME, must be systematically designed to utilize and enhance those factors which contribute to educational interchange. The second phase of the project will involve the design and implementation of such a systematic approach.

REFERENCES

1. Brown, C., Jr. and Uhl, H. Mandatory continuing education - sense or nonsense? JAMA 213:1660-1668, 1970
2. P.S.R.O. Program Manual. Rockville, Maryland: U.S. Department of Health, Education and Welfare, 1974.
3. The PEP Primer. Professional Evaluation Procedures for Auditing and Improving Patient Care. Chicago: Joint Commission on Accreditation of Hospitals, 1974.
4. Sandlow, L.J. Patient Care Evaluation: The Implementation of a Model System. Chicago: Michael Reese Medical Center, 1979.
5. Sanazaro, P.J. Medical audit, continuing medical education and quality assurance. West. J. Med. 125:241-252, September, 1976.
6. Peer Review Manual (2 vols.). Chicago: American Medical Association, 1972.
7. Ashbaugh, D.G. and McKean, R.S. Continuing medical education. JAMA 263:1485-1488, 1976.
8. Jensen, G., Liveright, A.A. and Hallenbeck, W. Adult Education: Outlines of an Emerging Field of University Study. Adult Education System of the U.S.A., 1964.

DEVELOPMENT AND EVALUATION OF A COMMUNITY-BASED PULMONARY EDUCATION SYSTEM

Roland G. Hiss, M.D., Jeoffrey K. Stross, M.D., Wayne K. Davis, Ph.D., and Roderick Macdonald, M.A., The University of Michigan Medical School

Traditional approaches to continuing medical education (CME) have several weaknesses. In general, CME courses are irrelevant to the practice setting of the individual physician, unstimulating due to the predominance of passive learning methods used, and inconvenient requiring the physician to travel and be absent from his or her practice. In small hospitals, the occasional educational event arranged for the hospital staff is too infrequent to have much impact. While medical journals and books are available in abundance, they generally require a large amount of reading to come up with a small amount of relevant information.

In this setting of inadequate and insufficient CME methods, the current effort is being made to develop new technology to deliver CME, especially for smaller hospitals. The method employs the informal communication network, which has heretofore not been deliberately cultivated as a CME mechanism. It is a pilot for a potentially much larger CME system.

PURPOSE:

The primary goals of this project are:

1) To develop, implement and evaluate a practical system of community based continuing medical education, delivered by physicians who have been identified by their peers as "educational influentials" (EI's).

2) To develop, test and perfect a set of pulmonary educational materials which can be utilized by any "educational influential" in a community setting.

3) To demonstrate the effectiveness of this educational system in improving primary care physician performance in the identification and treatment of chronic obstructive pulmonary disease.

BACKGROUND:

Within the informal communication network which exists in any social structure, a small community hospital for example, there are various forces which are responsible for the diffusion of new knowledge and skills throughout that social structure. One of these forces was first described by Lippet[1] as the "change agent," a helping individual who promoted change among his or her

---

This Project was supported by a contract with the Division of Lung Diseases, National Heart, Lung and Blood Institutes; Contract NO 1-HR6-2936.

Address reprint requests to: Dr. Hiss, Department of Postgraduate Medicine and Health Professions Education, University of Michigan Medical School, Ann Arbor, Michigan 48109.

colleagues. This concept was later developed by Havelock[2] who described the change agent as a catalyst, a process helper, or a resource lender. The existence of influential individuals on the staffs of small community hospitals was first identified by Wenrick, Mann, Morris and Riley,[3] in a few midwestern hospitals, and then in a total of 21 community hospitals[4] in Michigan ranging in size from 100 to 550 beds. These researchers confirmed the existence of what they termed "educational influentials" (EI's) in each hospital. These EI's functioned in the internal system primarily as information givers and obtained their own information from other sources, both inside and outside the system.

METHODS:

1. Community Hospital Selection: The target of this research project was small community hospitals. Two categories of hospitals were identified: "large" ones with 30-50 physicians on the staff, and "small" ones with 10-20 physicians. From a pool of hospitals in the State of Michigan which met these size criteria, and which had neither graduate medical education nor pulmonary-trained physicians, 16 hospitals were chosen on a first response basis to a general invitational letter. Each responding hospital was randomly assigned to either the experimental or control category, ultimately 8 in each. CME activities were instituted locally in each of the 16 hospitals as part of the project activities.

2. Educational Influential Identification: The method developed in this project for characterizing and identifying physicians as educational influentials has been previously reported.[5] The EI identification instrument was employed in the 16 project hospitals.

3. EI Recruitment: Following identification using the EI identification instrument, the EI's in the eight experimental communities were visited and the nature of the project described in detail. They were invited to participate, being asked to give their time and effort without compensation.

4. EI Training: A special and intense training program in chronic obstructive pulmonary diseases was developed specifically for the 8 educational influentials of this project. The major portion of the training program was offered in two four-day sessions about two months apart at the homesite of the project staff (Ann Arbor). Each session was attended by four EI's, the training program being offered twice. Pretraining materials were provided for study prior to the formal training period, and post-training support materials were given to the EI's to take home with them.

   A needs assessment to serve as a basis for the training program was accomplished first through a panel of primary care physicians from the State of Michigan who defined criterion care for COPD, and this was subsequently modified by consultants' review. Table I summarizes the instructional objectives of the program, both those for the EI, and those for the primary care physicians in his or her community who would ultimately be the recipient of informal training by the EI. Table II lists the major components of the training program offered the EI.

Table I. Instructional Objectives

Instructional objectives for the educational influential and for the primary care physician in the influential's home community were developed under the headings listed. For each heading, a knowledge and a skills category were developed with specific items under each written in behavioral terms.

I. Educational Influential:
   - A. Pulmonary Anatomy and Physiology
   - B. Pathophysiology of COPD
   - C. Clinical Signs and Symptoms of COPD
   - D. Management of COPD
   - E. Prevention of COPD
   - F. Function as Pulmonary Education Coordinator

II. Primary Care Physician:
   - A. Diagnosis of COPD
   - B. Primary Care Management of COPD
   - C. Referral of Patients with COPD
   - D. Prevention of COPD

Table II. EI Training Program

Pre-Training Material	Training Curriculum		Post-Training Support
1. Pre-Assessment Test - 35 multiple choice - Management, treatment, Dx 2. COPD Basic Reference - Pulmonary Medicine Guenter & Welch 3. Clinical Problems - Productive cough - Acute severe dyspnea - Progressive cough and dyspnea - Shortness of breath, strider 4. Blood Gas Manual	Interpretation of ABG's Clinical simulation Dx of COPD Pulmonary Function Tests Patient evaluation Bronchodilators, steroids IPPB Oxygen therapy Chest physiotherapy Small airways disease Upper airway obstruction	Complications of COPD Pulmonary defense mechanism Preoperative pulmonary evaluation Patient education Chronic bronchitis & emphysema Bronchiectasis Asthma	1. Post Assessment Test 2. Slide Tape Programs - McMaster University Series - Respiratory Physiology Series - Illustrated Lectures in Pulmonary Physiology 3. Review Session - Clinical Problems 4. Patient Education Slide Tapes

5. Educational Role of the EI: The EI's were selected because of their natural function within the informal communication network of their own hospital. This project deliberately intervened in this process by providing the EI with new knowledge and skills to employ during interactions with their primary care physician colleagues, primarily during the informal consultation process about individual patients. During this process, the primary care physician is in a position to receive continuing medical education which meets practically all of the ideal standards for a CME program--directly relevant to patient care, timely, convenient, practical, individually-based, and devoid of threat or embarrassment. The EI could capture the so-called "teachable moment" that arises around a clinical case.

6. Evaluation Strategies: Table III lists the multiple evaluation strategies employed in this project, and matches them against its three primary goals.

Inpatient and outpatient chart audits in the 8 experimental and 8 control communities were the principle methods. ICDA diagnostic codes relating to obstructive pulmonary disease were grouped, and all inpatient charts with these codes were reviewed for a 6 month period prior to the project intervention and for a comparable 6 month period post-intervention. This was accomplished in all 16 project hospitals. To determine the potential influence that a highly visible chart review might have on subsequent physician performance, the outpatient (office record) audit was performed using a Solomon Four design as shown on Table IV. For the outpatient review, the office records of COPD patients of primary care physicians in project communities were measured for the winter months of January and February before and after the educational intervention in 8 communities, and only after the educational intervention in 8 others.

Table III. Evaluation Strategies

Goals To Be Evaluated	Critique by E.I.'s	Patient Logs	Inpatient Chart Audit	Outpatient Chart Audit	Checklist of Facilities and Services	Consultant Review	Log of E.I. Contacts	Survey of Primary Care Physicians	Frequency of Dx and Rx Procedures
1. To develop and implement a practical system of community-based continuing medical education, etc.	X	X	X	X	X	X	X	X	X
2. To develop, test, and perfect a set of pulmonary educational materials, etc.	X					X	X	X	
3. To demonstrate the effectiveness of this educational system, etc.		X	X	X	X	X		X	X

Table IV. Experimental Design for Collection of Outpatient Audit Data

	Experimental Condition	Pre-Program Assessment	Treatment	Post-Program Assessment
Large Practice Communities (2 each)	1.	X	X	X
	2.	X		X
	3.		X	X
	4.			X
Small Practice Communities (2 each)	1.	X	X	X
	2.	X		X
	3.		X	X
	4.			X

7. Strategies of Analysis: Physician performance as measured by chart audit was analyzed using a scoring system, with weighted points being given for various components of performance. Measurable physician change in use of specific diagnostic and therapeutic maneuvers which had been included in the training program were noted in the experimental communities and compared to the performance of physicians in control communities on the same items.

RESULTS:

The educational influentials were identified successfully in 15 of the 16 project communities. Of these, 8 were invited to participate (the 7 in the control communities were not contacted in any way), and all 8 accepted the invitation with enthusiasm. All were trained according to the program outline and returned to their home community to apply this new knowledge and skills within the informal communication network which had identified them.

Preliminary analysis of initial data from the chart audit activities has shown some tendency for primary care physician behavior change in experimental communities, particularly in the prescribing of combination bronchodilator therapy. However, at the time of this writing, the major inpatient and outpatient chart audit is still under way, and definitive results are not available. It is anticipated that these will be available by August 1979.

DISCUSSION:

The need for new continuing medical education methodologies is clear, particularly techniques applicable to small community hospitals which do not have large graduate programs and bountiful human resources to sustain the traditional teaching hospital education program. The use of educational influentials selected from small communities using the methodology described by this project may demonstrate a means for providing useful local CME through an entirely different mechanism than the traditional means. If this system is perfected, it could be applied to a variety of other medical areas besides COPD, in which a single individual from a given community is selected for special intensive training (which would be cost efficient), and this individual could serve as a conduit for this training back to his or her colleagues at home.

The special nature of the educational influential is of considerable interest. These physicians command the respect of their colleagues as being good physicians, and in addition they possess valuable social communication skills and a personal energy that seems to spark more change than is the case for most of their colleagues. Further studies will be required, however, to demonstrate that only a physician with the special characteristics of an educational influential will be successful in the informal communication network. That other physicians in the community whose personal characteristics are not those of an EI could not also be effective has not as yet been shown. Similarly, we do not know whether the medical topic under consideration needs to be one generally managed by the discipline of the EI, or whether an EI can function "generically" and assist his or her colleagues in any medical area following special training. Answers to these questions will be the subjects of future investigations.

CONCLUSIONS AND IMPLICATIONS:

The educational influential may be a new method of continuing medical education with widespread applicability to a vast area of the practice community which now is beyond the perimeter of regular CME activities. It may be applied to a variety of other medical conditions and become a means for regularly updating small medical communities in a more effective, directly patient-related, way.

REFERENCES:

1. Lippet, R., Watson, J., Westley, B.: *The Dynamics of Planned Change*. Harcourt, Brace and World, Inc., New York, 1958.

2. Havelock, R.G.,: *The Change Agent's Guide to Innovation in Education*. Educational Technology Publications, New Jersey, 1973.

3. Wenrick, J.W., Mann, F.C., Morris, W. and Reilly, A.: Informal Education for Practicing Physicians, *J. of Med. Educ*. *46*: 4, April 1971.

4. Morris, W.: The Information Influential Physician: The Knowledge Flow Process Among Medical Practitioners. Unpublished Ph.D. Dissertation, The University of Michigan.

5. Hiss, R.G., Macdonald, R. and Davis, W.K.: Identification of Physician Educational Influentials (EI's) in Small Community Hospitals. Submitted for publication to the *Journal of Medical Education*.

PHYSICIAN PRACTICE PROFILES

R. Wayne Putnam, M.D., and Lynn Curry, Ph.D.
Dalhousie University, Faculty of Medicine

1. PURPOSE

Due to the currently growing emphasis on individualized education for physicians, we sought to determine (1) if Maritime Physician's practices were sufficiently different to warrant individual measurement; and (2) if there was a typical month that could be used on a general basis for individualized measurement.

2. BACKGROUND

For most of the early years of its development, Continuing Medical Education focused on group learning activities, usually "based on the intuitive assessment of need by teacher, programmer or student" (Meyer, 1973). Physicians who wished to participate attended refresher courses at a university or a scientific component of a professional society meeting. Although individual reading has always been a favoured way of learning, it has not been given equal status with formal programmes as "legitimate" CME, despite the lack of proof of the usefulness of the latter (Lewis and Hassanein, 1970; Bertram and Brooks-Bertram, 1977).

However, in the past decade, much more attention has been given to individualizing practitioners' educational needs. Examples are the Individual Physician Profile developed in Wisconsin (Sivertson, 1973; Sivertson, 1974), the multitude of self evaluation and self assessment programmes produced by specialty colleges, and the Practise Related Education Program (Bowler, 1977). Patient care appraisal in hospitals and clinics has attempted to relate CME to patient care given by groups of physicians with the potential for individual feedback regarding inadequacies of care delivered (Fleisher, etal 1976; Brown, Fleisher, 1971; Davidson, etal 1974; Hanson, Kraus, 1973).

Much of the attention in the CME literature recently has been about mandatory Continuing Medical Education and its relationship to relicensure or recertification (Brown, Uhl, 1970; Richards, 1976; Stross, Harlan, 1978). In this application also, the emphasis has shifted from group learning to individualized education.

---

* This paper is based on research funded by Health and Welfare Canada, NHRDP grant #6603-1075-44

** Requests for reprints should be sent to R. Wayne Putnam, M.D., Director, Division of Continuing Medical Education, Dalhousie University, Halifax, Nova Scotia, Canada. B3H 4H7

Mueller (1978), has concluded that, "In continuing evaluation, the evaluation will not be comprehensive for any specialty as defined by either educators or evaluators but will be relatively unique to the individual practitioner; it will be practice profile oriented; carefully tailored but not necessarily totally individual;..."

As this trend continues, it becomes necessary for continuing education programs to justify the expense of individual measurement and planning by careful evaluation of their target audience. In order to most efficiently supply this service we need to determine where the variation exists, and how it is patterned by time.

3. METHODOLOGY

The research design described is predicated upon another study which was open to all family physicians in the Maritime Provinces of Canada who had been in practice from five to 25 years and who had access to hospital beds. From the positive responses to an invitation letter, a sample of 16 family physicians were randomly selected, representing both urban and rural practices, solo and group settings. No two physicians were from the same community. Each physician kept a list of every diagnosis managed in the office for two separate six month periods (July to December '77, and January to June '79). The first time block was available for this analysis. Multiple diagnoses were included if each one received attention during the visit. All diagnoses were coded by the International Classification of Health Problems on Primary Care and were entered into a computer for storage and eventual production of profiles.

Diagnostic codes were compiled by month for each physician, yielding numeric profiles of the monthly pattern of practice. Chi square analyses were conducted to examine the variability of months and individual physicians. The pattern of chi squares was tested for significant differences by analysis of variance, and the interaction term estimated by a Tukey's t-test.

4. RESULTS AND DISCUSSION

Table 1. contains the chi squares comparing each month for each physician to his overall profile (Figure 1.). No immediate pattern of differences was obvious, so the analysis of variance presented in Table 2. was conducted. Significant differences ($F = 2.49$, $p = .004$) are noted between doctors, but not between months ($F = 1.59$, $p = .160$). The estimated interaction term is not significant (sum of square = .314) indicating that the main effects operate independently. This means that no one month better approximates an individual's six month profile.

	July 1977 #1	Aug. 1977 #2	Sept. 1977 #3	Oct. 1977 #4	Nov. 1977 #5	Dec. 1977 #6	Jan. 1978 #7
DOCTOR 1	15.16	15.67	17.15	15.27	12.13	12.96	
DOCTOR 2	14.20		13.62	19.57	17.30	13.17	11.65
DOCTOR 3	15.65	17.47	19.64	15.70	18.98		22.15
DOCTOR 4	13.92	18.10	19.30	16.35	16.54		12.56
DOCTOR 5		17.38	17.86	19.28	19.91	15.01	15.79
DOCTOR 6		18.56	16.16	16.09	18.27	12.20	10.55
DOCTOR 7	17.27	18.62		19.74	17.51	16.50	18.71
DOCTOR 8		16.98	15.92	12.30	17.73	17.67	10.35
DOCTOR 9	15.99		15.45	18.73	14.50	12.95	19.41
DOCTOR 10		15.88	17.05	15.27	17.30	16.59	22.37
DOCTOR 11	13.88	15.55	14.66	16.91	16.06		17.91
DOCTOR 12	15.12		21.25	18.74	16.68	13.72	22.94
DOCTOR 13				15.25	16.42	15.71	14.49
DOCTOR 14		12.52	16.00	17.00	18.24	16.55	18.63
DOCTOR 15		20.26	16.99	16.41	19.90	14.36	19.68
DOCTOR 16		22.93	19.84	19.39	18.66	19.17	12.08

TABLE 1.

Chi Square Analysis: Doctors' profile and month profile compared to group profile

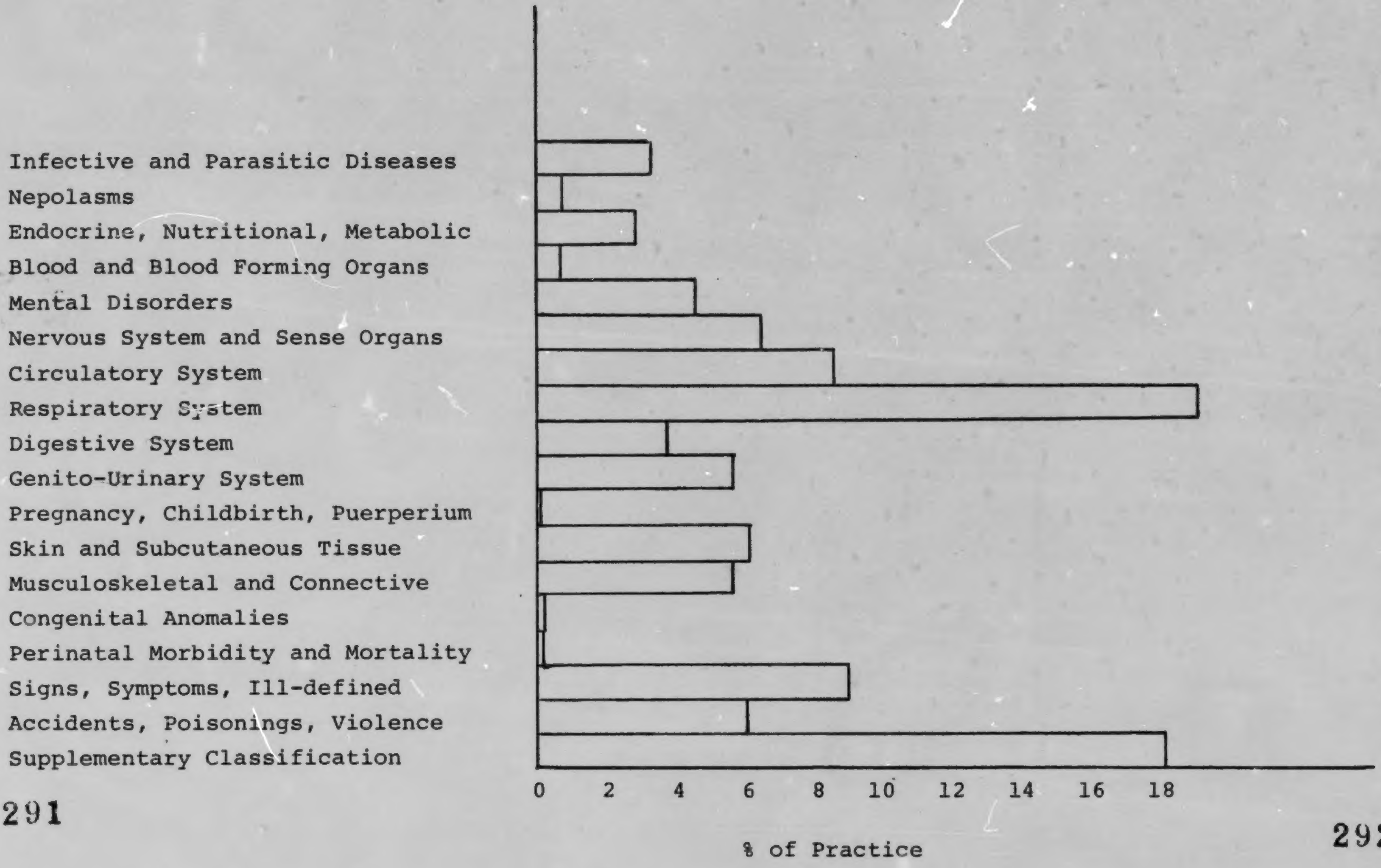
Source of Variation	Sum of Squares	DF	Mean Square	F	Signif of F.
Main effects	238.573	21	11.361	2.236	.005
Doctor	190.191	15	12.679	2.496	.004
Month	48.382	6	8.064	1.587	.160
Explained	238.573	21	11.361	2.236	.005
Residual	457.202	90	5.080		
Total	695.775	111	6.268		

112 cases were processed.

0 cases (0PCT)were missing

Table 2: Analysis of Variance

Figure 1: Group Profile (Average across doctors)



5. CONCLUSIONS AND IMPLICATIONS

It is clear from this analysis that there are important differences in the practice patterns of Maritime physicians. As we try to comply with trend to personalize educational assessment and planning, we will be forced to measure our participants for individual differences in practice patterns. These differences in our target population are real, so the product and the process offered to them should be expected to reflect these differences. This is a more data based reason for personalized assessment than that offered by Sivertson (1973) where he asked participants to judge the relevance of various topics to their practices.

This analysis informs a second point important to the operationalization of personalized educational assessment and planning. Since there is no significant differences across months for any physician, it would be sufficiently accurate to take any one month of diagnostic codes as a sample.

In fact, the question then becomes: What is the minimum time sample that still accurately reflects the total profile? Several different methods are currently being used to obtain these profiles, including listing 100 consecutive episodes of care (Bowler, 1977), 150 and 300 consecutive episodes of care (Rainsberry, 1970) and listing all the episodes of care managed on a different day each week for four weeks (Sivertson, 1974). These methods have not been validated by comparison with total profiles, such as made available by this study.

After replication of the reported findings with the second time block, our group will undertake this analysis to discover the minimal accurate time sample to profile physician practices. This determination will be important to planners of personalized educational assessment and planning.

References

Bertram, D.A. & Brooks-Bertram, P.A. The evaluation of continuing medical education: a literature review. *Health Education Monographs*, 1977, *5* (4), 330-362.

Bowler, F.L., Brading, P.L., Burg, F.D., Finestone, A.J., and Hubbard, J.P. A practical related educational program. *Journal of the American Medical Association*, 1977, *237* (13), 1346-1349.

Brown, C.R., amd Fleisher, D.S. The bi-cycle concept - relating continuing education directly to patient care. *New England Journal of Medicine*, 1971, *284* (Suppl.), 88-97.

Brown, C.R., and Uhl, H.S.M. Mandatory continuing education: sense or nonsense? *Journal of the American Medical Association*, 1970, *213* (10).

Davidson, R.C., Lein, J.N., and Kelday, J.M. Quality assurance in the provision of hospital care, Case Study, Audit network aids continuing medical education. *Hospitals*. 1974, *48*, 85-88.

Fleisher, D. S., Brown, C. R., Zeleznik, C., Escovitz, G. H., & Omdal, C. The mandate project: institutionalizing a system of patient care quality assurance. *Pediatrics*. 1976, *57* (5), 775-782.

Hanson, A. S., Kraus, E. D. An outpatient medical audit. *Minnesota Medicine*. 1973, *56* (Suppl. 2), 49-52.

Lewis, C. E., & Hassanein, R. S. Continuing medical education - an epidemiologic evaluation. *New England Journal of Medicine*. 1970, *282* (5), 254-259.

Meyer, T. C. Toward a continuum in medical education. *Journal of Medical Education*. 1973, 48 (Suppl.), 67-70.

Mueller, Barber *Federation Bulletin*. 1978, 65 (12).

Rainsberry, Paul, of College of Family Physicians of Canada. Personal Communication.

Richards, R. K. Mandatory CME - springboard of stumbling block? *Michigan Medicine*. 1976.

Sivertson, S. E., Hansen, R. H., Shropshire, R. W., & Schonenberger, A. O. Family practices in Wisconsin: implications for medical education and delivery of health care. *Wisconsin Medical Journal*. 1974, *73*, 170-174.

Sivertson, S. E., Meyer, T. C., Hansen, R., & Schoenberger, A. Individual physician profile: continuing education related to medical practice. *Journal of Medical Education*. 1973, *48*, 1006-1012.

Stross, J. K., Harlan, W. R. The impact of mandatory continuing medical education. *Journal of the American Medical Association*. 1978, *239* (25), 2663-2666.

PRECIS

STUDENT SATISFACTION AND STRESS

MODERATOR: C. Warner Johnson, M.D.
University of Southern California

IDENTIFYING SATISFYING AND STRESSFUL EVENTS IN MEDICAL SCHOOL

A method for identifying everyday events which contribute to the positive or negative affective state of medical students is presented. Correlation and regression analysis is used to investigate the relationship of academic and non-academic events to medical student mood.

MEDICAL VS. GRADUATE STUDENTS: SOCIALIZATION AND STRESS

A study comparing medical and graduate students to compare issues of socialization and stress. Conclusions drawn that socialization is later and stress greater for medical students.

DOGMATISM AND VALUES ORIENTATION OF RESIDENTS IN FAMILY PRACTICE

This study compared first-year residents in family practice with first-year residents in two consultant specialties on a dogmatism scale and a value survey. Family practice residents valued helpful and true friendship more than the consultant specialists did; however, there was no difference in mean dogmatism scores.

Identifying Satisfying and Stressful Events in Medical School

Richard Maisiak, Ph.D., James Britton, III, and
Ronald Meredith, Psy.D.
University of Alabama in Birmingham

Literature Review

Past research (1-6) has shown that medical students undergo considerable stress during their medical education. Kilpatrick, Dubin, and Marcotte (3) have reported that medical students undergo increased mood disturbances and exhibit a high level of anger and hostility especially in their first year. Boyle and Coombs (4) found that fears of inability to absorb material, fears of getting bad grades, and a limited social life were all among the chief sources of stress for freshman medical students. Edwards and Zimet (5) found that lack of personal freedom, excessive academic pressures, and feelings of dehumanization were among the chief concerns of medical students. Student anxiety can have detrimental effects in many areas. Barrett and White (6) found that highly anxious students were more likely to have a decreased academic performance and are more likely to seek psychiatric help. Researchers in other fields have shown a strong relationship between stress, depression, and illness and stress and depression have been shown to be a product of unpleasant events (7-11).

Purpose

The general objective of the present research was to develop a method for identifying the life events of medical students which contribute most to stress and to satisfaction. A specific objective was to investigate the influence of different classes of life events on both student happiness and unhappiness. If specific, concrete, stressful and satisfying events could be identified, then intervention strategies could be employed to curb negative events and promote positive ones. The present research plan differs from past studies in that the present plan involves the direct measurement of student mood and seeks to identify events contributing to student satisfaction as well as stress. The main hypothesis is that certain concrete events can be identified which are significantly related to medical students' affective states. The present study investigates, through correlation and regression analysis, the relationship between students' self-reported life events and their positive and negative mood states as measured by a valid test.

Method

Subjects The subjects were all one-hundred and sixty-eight students of the 1978-1979 freshman medical school class at a large medical school. Most students were male and aged between twenty-one and twenty-nine. One-hundred and six (63%) students completed the Medical Student Events questionnaire upon request.

Instrument The Medical Student Events (MSE) questionnaire consisted of the Bradburn Affect Balance Scale, a list of possible medical student life events, and a biographical section. The Bradburn Affect Balance Scale (12, 13) is a ten-item yes-no scale developed by Bradburn to assess psychological well-being. The scale is designed to measure positive affect (analogous to happiness), negative affect (analogous to unhappiness or depression), and total affect, the balance of positive and negative affect. Scores on all three scales range from 0 to 5. A higher score represents either higher positive affect, higher negative affect, or higher total affect. The scale has been used in a study of the relationship between national events and national mood. The scale has been shown to be reliable and valid (12). The negative affect scale has been shown to be positive correlated with measures of depression. In the present study subjects were asked to respond in the way they have felt in the previous two weeks.

The medical events section of the MSE consisted of a list of 36 life events in five categories including academic events, social events, financial events, health and exercise events, and recreational events. Specific events in each category were constructed after discussions with medical students. Events were selected that could occur in everyday life. The final section of the Medical Student Events questionnaire were items concerning the sex, marital status, and age of the respondent.

Procedure Students received the MSE questionnaire in their student mailboxes. The students were told that their responses were anonymous and confidential and that they could refuse to complete any item of the MSE. Few non-responses occurred. Students were asked to recall events that occurred in the previous two-week period.

Statistical Analysis Student responses were analyzed in a two-part statistical procedure. In the first analysis, the self-reported occurrence of any life event was coded as "1" and the non-occurrence of any life event as "0". Correlations were then computed between each of the three Bradburn Affect Scale scores and each of the medical student events. Correlations were also computed between the affect scale scores and the biographical data. In the second analysis, three stepwise multiple regression analyses were performed between the three affect scale scores, the dependent variables, and all the life events, the independent variables. The process was used to identify the medical student life events most predictive of student mood. Although there are some statistical problems in correlation and regression analyses involving dichotomous variables, the general conclusions of such analyses are unlikely to be altered by these problems (see Neter and Maynes, 14).

Results and Discussion

The correlations of various medical student academic events with positive, negative, and total affect scale scores are presented in Table 1. In the present study, any correlation above .25 or below -.25 is statistically significant (p <.01). Certain events which would be expected to be related to a satisfying mood were positively correlated with the Positive Affect Scale scores. "Receiving a 'good' grade", "preparing well for an exam", and "successfully memorizing difficult study material" were positive correlated with the Positive Affect Scale score and Total Affect Scale scores. Surprisingly, "assisted another student in learning" and "having a good discussion with another student" produced even higher correlations with the Positive Affect Scale and the Total Affect Scale. No academic events produced a strong correlation with the Negative Affect Scale scores. The implications are that the academic events have a stronger influence on positive affect than on negative affect. Academic events involving student interaction were especially related to the Total Affect score.

The correlations of the medical student social and financial events with Positive, Negative, and Total Affect Scale scores are presented in Table 2. By far the social event producing the highest correlation with the Positive, Negative, and Total Affect Scale scores is "making new friends." Bradburn (12) also found that this event was related to affect, especially with the Positive Affect Scale. Interestingly, "attending parties", "sexual intimacy", and "club participation" were events that produced low correlations with the Affect Scale scores. The financial event, "being short of money", produced a positive correlation with Negative Affect. Other financial events were unrelated to the Affect Scale scores.

The correlations of medical student health and recreation events with Positive, Negative, and Total Affect Scale scores are presented in Table 3. The only health and exercise event producing a significant correlation with the Total Affect Scale was "having trouble getting up in the mornings". Other health events produced low correlations with the Affect scales. The recreational event which produced the highest correlation with Total Affect was "going outdoors for recreation". Interestingly, "going outdoors" produced a significant positive correlation with the Positive Affect Scale score but a low correlation with Negative Affect. Other recreational events produced lower correlations. Correlations of the biographical data will affect were low.

The preceding analysis examined the relationship of classes of events with affect. A stepwise regression analysis was performed to identify those individual events which were most predictive of the Total Affect scores. The four events which contributed the most to the regression equation are in Table 4. Two of the events, "had a good discussion with a student" and "assisted another student in learning" were academic events. When taken together all four events contributed to 36.4% of the variance of the Affect scores.

Conclusions

The main finding of the present study was that the concrete events in medical students' lives which have influence on their affective states could be identified. The present method involved a direct measurement of mood and a direct assessment of the extent of the influence of different events on both positive affect, a satisfying state, and negative affect, a potentially stressful state. Using the present method educators can select out the proper events to promote and to eliminate in the medical curriculum. Marshall (15) has suggested that the measurement of the aspects of the learning environment relevant to student stress is a necessary step prior to planning appropriate interventions. The specific academic events mentioned in the present study that are most likely to contribute to increasing the satisfaction and reducing the stress of freshman medical students are "assisting another student in learning" and "having a good discussion with another student." Educators might consider encouraging more group study, tutoring, or group discussions in the curriculum. In the present study no attempt was made to list all possible medical school events. Further research might be directed to assessing the affects of other events on student satisfaction and stress.

Table 1

Correlations of Various Medical Student Academic Events With Positive, Negative, and Total Affect

	Positive Affect	Negative Affect	Total Affect
Assisted another student in learning	.30	-.18	.34
Had a good discussion with another student	.30	-.18	.33
Prepared well for an exam	.28	-.16	.31
Received a "good" grade	.22	-.07	.21
Attended an exciting lecture	.23	-.06	.21
Solved a difficult academic problem	.15	-.08	.17
Successfully memorized difficult study material	.29	.07	.17
Enjoyed a good lab experience	.14	-.07	.15
Helped a patient or friend with a medical problem	.22	.03	.14
Received a "bad" grade	.01	.18	-.11
Had a good conversation with a faculty member	.02	-.06	.06

Table 2

Correlations of Medical Student Social and Financial Events with Positive, Negative, and Total Affect

	Positive Affect	Negative Affect	Total Affect
Social Events			
Made any new friends	.26	-.22	.33
Attended a party	.16	-.12	.19
Solved a difficult personal problem	.19	-.04	.17
Became less friendly with a friend	.05	.17	-.15
Had satisfying sexual intimacy with another	.18	-.03	.15
Attended church	.03	-.14	.12
Talked with parents, relatives, or siblings	.15	-.03	.13
Participated in an organized activity (intramural, club)	-.02	.11	-.08
Financial Events			
Been short of money	.02	.25	-.15
Received a paycheck	-.02	-.13	.06
Made a savings account deposit	.02	.05	-.02
Made a checking account deposit	-.04	-.01	-.02

Table 3

Correlations of Medical Student Health and Recreation Events with Positive, Negative, and Total Affect

	Positive Affect	Negative Affect	Total Affect
Health and Exercise			
Had trouble getting up in the mornings	-.12	.27	-.27
Felt very exhausted	-.07	.14	-.15
Had a good physical workout	.12	-.04	.11
Had a cold or flu	-.12	.03	-.11
Had trouble getting to sleep	-.03	.11	-.10
Had an illness other than cold or flu	.03	.07	-.02
Recreation			
Gone outdoors for recreation	.24	-.07	.22
Eaten an especially good meal	.10	-.16	.18
Taken an overnight trip	.10	-.10	.14
Started a new hobby or pastime	.10	-.06	.12
Done something fun for the first time	.09	-.04	.10
Consumed liquor heavily	.05	.00	.04
Gone out to eat	-.11	-.14	.01

Table 4

Percentage of Total Variance of Total Affect Accounted for By Four Medical Student Events in a Multiple Regression Analysis

	Percentage of Variance
Had a good discussion with another student	11.4%
Had trouble getting up in the mornings	14.0%
Assisted another student in learning	5.0%
Eaten an especially good meal	6.0%
TOTAL	36.4%

References

1. Adsett, C. A. Psychological health of medical students in relation to the medical education process. *Journal of Medical Education*, 1968, *43*, 728-734.

2. Rosenberg, P. P. Students' perceptions and concerns during their first year in medical school. *Journal of Medical Education*, 1971, *46*, 221-218.

3. Kilpatrick, D. G., Dubin, W. R. and Marcotte, D. B. Personality, stress of the medical education process, and changes in affective mood state. *Psychological Reports*, 1974, *34*, 1215-1223.

4. Boyle, B. P. and Coombs, R. H. Personality profiles related to emotional stress in the initial year of medical training. *Journal of Medical Education*, 1971, *46*, 882-887.

5. Edwards, M. T. and Zimet, C. N. Problems and concerns among medical students, 1975. *Journal of Medical Education*, 1976, *51*, 619-625.

6. Barratt, E. S. and White, R. Impulsiveness and anxiety related to medical students' performance and attitudes. *Journal of Medical Education*, 1969, *44*, 604-607.

7. Vinokur, A. and Selzer, M. L. Desirable versus undesirable life events: Their relationship to stress and mental illness. *Journal of Personality and Social Psychology*, 1975, *32*, 392-337.

8. Dohrenwend, B. S. and Dohrenwend, B. S. (Eds.), *Stressful life events: Their nature and effects*. New York: John Wiley and Sons, 1974.

9. Lewinsohn, P. M. and Graf, M. Pleasant activities and depression. *Journal of Consulting and Clinical Psychology*, 1973, *41*, 261-268.

10. Lewinsohn, P. M. and Talkington, J. Studies on the measurement of unpleasant events and relations with depression. *Applied Psychological Measurement*, 1979, *3*, 1, 83-101.

11. Rabkin, J. G. and Struening, E. L. Life events, stress, and illness. *Science*, 1976, *194*, 1013-1020.

12. Bradburn, N. *The structure of psychological well-being*. Chicago: Aldine, 1969.

13. Bradburn, N. and Caplovitz, D. *Reports on happiness*. Chicago: Aldine, 1965.

14. Neter, J. and Maynes, E. S. On the appropriateness of the correlation coefficient with 0, 1 dependent variables. *Journal of the American Statistical Association*, 1970, *65*, 501-59.

15. Marshall, R. E. Measuring the medical school learning environment. *Journal of Medical Education*, 1978, *53*, 98-104.

MEDICAL vs. GRADUATE STUDENTS: SOCIALIZATION AND STRESS

BY: Joanna Spiro, Ed.D., Marcella Roenneburg, and Betty Joan Maly, M.D.
The Medical College of Wisconsin

Medical educators have shown increasing interest in the stress patterns and general adaptation of medical students.[1-4] Complaints of lack of personal freedom, excessive academic pressures, feelings of dehumanization, loneliness, and insecurity are ascribed to the demands of a medical education.[2,5]

Among the frustrations of researchers attempting to evaluate these reports is the absence of studies comparing medical students to other students at the post-bacculaureate level. Such comparisons are necessary if we are ever to determine whether the reports of medical students about the stresses encountered during their training are indeed greater than those who are pursuing other courses of advanced study.

This work begins with the premise that it may be useful to establish both commonalities and differences between medical students and other graduate students. The populations compared in this paper include medical students matched to a randomly selected group of graduate students at a univesity with which the medical school had once been affiliated.

Comparisons of these two populations included general demographic data such as marital status, living arrangements, religion, birth order, and parental occupation, as well as indicators of stress and anxiety. The latter included weight changes, sleep patterns, alcohol and drug use, interpersonal relationships, satisfaction with choice of career, and perceived support and understanding of career demands and difficulties by parents and close friends.

METHODS

The populations studied were surveyed by means of an anonymous questionnaire containing 36 questions. Each question was followed by a series of choices. Recipients were instructed to check all appropriate answers for each question. The questions were designed with two major foci. The first consisted of questions designed to obtain background and demographic information which would allow us to compare the two populations. The second was comprised of questions to evaluate stress and coping mechanisms within each group. Typical questions are demonstrated in the tables of results.

The names and addresses of all 1,139 graduate students registered at a nearby university in the spring of 1978 were obtained from the registrar's office. Questionnaires were mailed to a random sample of 950 of these graduate students. Over a three month period 46% were returned (N=441).

All junior and senior medical students received the questionnaires by mail. Freshman and sophomore surveys were distributed during lecture periods. Of the

---

The Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, Wisconsin 53226.

total 544 medical students, 82% responded (N=444). Questionnaires were either returned by mail or at designated drop off boxes.

Each questionnaire returned was coded and tabulated by computer. Total responses and percentages were established for each answer to every question. Chi-square analysis was performed to establish the significance of comparisons with $p < .05$ designated as the limit of acceptable significance.

RESULTS

In the demographic data similarities occurred in the numbers living alone, in birth order, in paternal occupation, and in parental income.

Differences in the demographic data were that twice as many graduate students as medical students were married or divorced. Although involvement with a significant other was greater among medical students it did not account for the disparity in marital status between the two populations. Correlating with this, twice as many graudate students lived with a significant other. Conversely, the number of medical students living with their parents or with roommates was much greater than the number of graduate students in the same situation.

Religious distribution by birth showed significantly more graduate students identifying themselves as Catholic, while medical students identified themselves significantly greater in Protestant and Jewish categories. The differences in religion by choice show a sharp decrease in affiliation with Catholicism for medical students, little change between the groups in choosing Protestantism, and a significant shift among medical students to categories of atheist, agnostic, and other. The Jewish percentage by choice remained approximately the same as it had by birth.

Differences in maternal occupation included more homemakers and skilled laborers among medical student mothers. More of the graduate students had parents who were retired.

The patterns of these two populations in alcohol and drug usage were similar in the categories of social drinking, drinking to excess, reasons for using illegal drugs, and the types of illegal drugs used.

Differences were found in tobacco and alcohol use in that twice as many graduate students smoked and reported drinking every day as did medical students. Twice as many medical students, however, reported using illegal drugs.

Emotional preoccupations of both groups involved equivalent concerns about overcommitment and appropriate career choice. There were no significant differences in reported weight change patterns, sleeping patterns and the number of depressive episodes since matriculation.

Graduate students had significantly greater concerns about the way they function sexually and their health problems as related to anxiety. The concerns of medical students dominated the categories of competency, making mistakes, the quality of relationships with a significant other, and prospects of ever having an intimate relationship. Although there were no differences in the number of depressive episodes and the number of students having counseling before entering postgraduate work, significantly more medical students reported that they had counseling since medical school started.

There were no differences in sexual preference between the two groups. However, masturbatory activity was more prevalent among medical students. Sexual activity with a partner among medical students had a bimodal distribution in that their frequency was either never/rarely, or three or more times per week. In contrast, the graduate students' data showed a skewed Gaussian distribution, the frequency of sexual activity with a partner being greatest at once a week or less.

Neither group felt that interpersonal relationships were enhanced during this stage of their education. To a significant degree, however, medical students more than graduate students felt that their personal relationships had suffered during the process of their education.

Medical students and graduate students did not appear to be different in their assessment of the understanding that significant others and friends had for the time, energy, commitment and strain involved in the student's education. The difference was that parental understanding in these areas was reported as being significantly higher by the medical students.

DISCUSSION

Some of the similarities of our two populations led us to believe that we were comparing students who came from families of approximately the same socioeconomic background. Parental occupations were remarkably alike as were their yearly incomes. Both groups were dominated by the first born members of families, which concurs with an earlier study on a medical student population.[6]

Although the age of the respondents was not part of our data base, it could be inferred that the graduate student population was older because a greater proportion of them were married and because more of their parents were retired. Age could also account for the fewer virginal graduate students.

There were differences in religion (by birth), but some of that difference may be attributed to the fact that the graduate student population attended a Catholic university. We were therefore not surprised by the disproportionately low number of Jewish graduate students in our study, nor at the larger number of Catholic graduate students. It is important to note that the medical school studied was, until recently, affiliated with the same university from which the graduate student population was selected. At the time of this study, both schools were located on the same campus, necessitating shared health, dining, library, and recreational facilities.

One problem area identified in the results is the concern medical students have about developing interpersonal relationships. Previous works indicates this, along with loneliness and a sense of isolation, as high ranking concerns of medical students.[1,3,7] Though they may progress as a large group through their education, their cohorts may be an unreliable source of support.[8] At the same time, medical students in general have less available time for developing relationships outside their study groups.[2,5] Thus, the emotional preoccupation with the quality of significant other relationships, concern over ever having an intimate relationship, and the feelings that their personal relationships had suffered during medical school become a highly characteristic profile of medical students.

304

To focus on their socialization process, one finds that 65% of the medical students are single. Only 7% identify themselves as related to a significant other. Thus, with the low numbers of married and divorced students, one sees a marked contrast between medical and graduate students. Although these differences may be attributable to age differences in the two populations, it represents a difference in support mechanisms available to medical students. This is particularly noticable in scrutinizing data on living arrangements in that less than half of the medical students who are in relationships with significant others live with them. 46% of medical students work out their living needs by living with parents or roommates - more than twice the graduate students in these categories. One questions whether these data mean that medical students are delayed in their formation of more socially mature relationships in which one makes a choice of a partner with whom one develops in life? Or is it merely that they are chronologically younger? If there is a delay in social development, is it directly related to lack of time and financial problems? Perhaps the choice of a life partner is less of a priority for medical students, or they may even feel hesitant to impose the process of their education upon a steady companion.

Medical students did reveal that more of their parents were understanding of the pressures and commitment involved in their education. This would be a logical consequence of the well defined role of the physician within our culture. Yet one cannot infer that medical students derive more support from this understanding, since it may only make it easier for a student to live in the parental home. Moreover, both parents and friends may contribute to isolation of medical students by helping them avoid social demands that might compete with academic involvement.

Medical students may even impose upon themselves another form of isolation in that 33% considered themselves estranged from the religious affiliation of their birth as compared with a 20% shift among graduate students. Most of them identify themselves as agnostic or atheist rather than choosing another faith. In our data base we cannot identify factors in medical education that may account for the difference between the two populations, though it is possible that it is an artifact in comparing medical students with a Catholic graduate school. The loss of religious identity, however, amy be another indicator of the narrow support base and altered socialization with which a medical student completes an education.

Limited time for socializing may be refelcted in fewer medical students reporting that they drink alcohol everyday, especially since they were similar to graduate students in incidence of social and excessive drinking. Another reason for the lesser frequency of alcohol use among medical students could be their recognition of the pathology implied in daily use.

Patterns of sexual activity further delineate socialization problems in that masturbatory activity is much more prevalent among the medical students in this study, and also in earlier work comparing medical students to the Kinsey report.[9,10] Mudd and Siegel attributed the increase frequency of masturbation to stress and anxiety. Yet Gotheil and Freedman found that sexual intercourse was less frequently an outlet for medical students than in a matched population of the Kinsey report, even though medical students tended to acquire their first experiences at an earlier age than Kinsey's population and were similar in other parameters such as duration of intromission. The data in our study also show that medical students presented the extremes in frequency of sexual activity with a partner, while graduate students showed a greater tendency to sexual activity once a week. The large number of married graduate students may account for this, but the unpredictable social availability of medical students may also be a factor, such that when the opportunity is present, medical students make the most of it.

If, as we have pointed out, medical students are different in their socialization patterns, then they are also likely to be different in coping mechanisms used in response to stress. The similarities in the two populations' concerns about over-commitment, appropriate career choice, weight change, sleeping patterns, and the number of depressive episodes may be indicators that both groups are in equilibrium with the stress they are experiencing although the levels of stress may be different. Among the indicators that medical students are experiencing more stress is the significantly greater number of medical students who had sought counseling since matriculation even though both populations had access to free counseling.

Indeed issues of competency and making mistakes were of greater concern for medical students in our study than health issues, as in other works[2,5], despite studies documenting a high prevalence of nosophobia.[11,12] For the third and fourth year medical students issues of competency and making mistakes are an everyday reality. These findings concur with those of Rosenberg who concluded that "the violence done to the self-image of the medical student is the most serious consequence of his educational program".[13]

Differences in coping behavior are demonstrated by greater tobacco use among medical students. It is understandable that medical students are less likely to use tobacco because of the strong medical arguments against its use and their exposure to the consequences of such use. Although the type of illegal drugs used and the reason for using them were similar in both populations, more than twice as many medical students reported using illegal drugs. Other researchers have recorded comparable figures.[14] This extensive drug usage may appear paradoxical considering their medical knowledge, but their knowledge of drugs and drug effects as well as their exposure to the use of these drugs may actually contribute significantly to the tendency to experiment by eliminating the element of fear and by knowledge of desired effects.[15] It is important to note that this degree of drug usage may correlate with high levels of stress experienced with few alternative releases or more acceptable coping mechanisms. This usage further emphasizes the poor socialization patterns, lack of personal time, loneliness, and isolation the medical student experiences.

BIBLIOGRAPHY

1. Boyle, B.P. & Coombs, R.H., *Personality Profiles Related to Emotional Stress in the Initial Year of Medical Training*, Journal of Medical Education, 46: 882-887, 1971.
2. Edwards, M.T. & Zimet, C.N., *Problems and Concerns Among Medical Students -- 1975*, Journal of Medical Education 5: 619-625, 1976.
3. Golden, J.S. & Rosen, A.C., *Group Dynamics for Medical Students*, International Journal of Group Psychotherapy 25:, 303-313, 1975.
4. Thomas, C.B., *What Becomes of Medical Students: The Dark Side*, Johns Hopkins Medical Journal 138: 185-195, 1976.
5. Coburn, D. & Jovarsas, A., *Perceived Sources of Stress Among First-Year Medical Students*, Journal of Medical Education 50: 589-595, 1975.
6. Cobb, S. & French, J.R.P., *Birth Order Among Medical Students*, JAMA 195: 172-173, 1966.
7. Spiro, J.H., Roenneburg, M., Maly, B.J., *Teaching Doctors to Treat Doctors: Medical Peer Counseling*, Journal of Medical Education, 53: 997, 1978.
8. Snow, T., *Preliminary Observations of the Psychotherapy of Medical Students*, American Journal of Psychotherapy, 23: 293-302, 1969.
9. Mudd, J.W. & Siegel, R.J., *Sexuality-Experience and Anxieties of Medical Students*, NEJM, 281: 1397-1403, 1969.
10. Gottheil, E. & Freedman, A., *Sexual Beliefs and Behavior of Single, Male Medical Students*, JAMA 212: 1327-1332, 1970.
11. Mechanic, D., *Social Psychologic Factors Affecting the Presentation of Bodily Complaints*, NEJM 286: 1133, 1972.
12. Hunter, R.C.A., Lohrenz, J.G., & Schwartzman, A.E., *Nosophobia and Hypochondriasis in Medical Students*, Jounral of Nervous & Mental Diseases 139: 152, 1964.
13. Rosenberg, P.O., *Student's Perceptions and Concerns During their First Year in Medical School*, Journal of Medical Education 46: 211-218, 1971.
14. Blaine, J., Tieherman, C., & Hirsh, J., *Preliminary Observations on Patterns of Drug Consumption Among Medical Students*, International Journal of Addictions, 3: 389-396, 1968.

Dogmatism and Values Orientation of Residents in Family Practice

Kenneth J. Wade, Ph.D.
Educational Evaluator: Office of Educational Development
School of Medicine
University of Alabama in Birmingham

Roland P. Ficken, Ph.D.
Chairman, Department of Behavioral Science
College of Community Health Science
University of Alabama

INTRODUCTION

The twentieth medical specialty became official on February 8, 1969, with the establishment of the American Board of Family Practice. Within eight years, the number of residency programs grew from none to 325, the number of residency positions grew from none to 5,421. During this same time, more than 80% of all U.S. medical schools developed programs in family practice (Willard, 1978). A major factor in the growth of family practice training has been federal funding policies and regulations. In general, the federal government has sought to increase the number of primary care physicians by its medical education funding policies.

As training of family practice physicians has expanded rapidly in the past 10 years, the question of personality orientation has become relevant, i.e., what type of personality best fits primary care? The matter of personality orientation is important because of the very nature of family practice and the need to fit physicians with roles which will be satisfying. Recent studies (Holden & Levit, 1978; Wechsler, Dorsey & Bovey, 1978) have reported data showing a significant proportion of physicians who train in primary care later seek another type of medical practice. This would suggest that many who train for the role of primary care practitioner later find that role to be incompatible. The plans of government policymakers seeking to increase the supply of primary care practitioners may be hindered by a significant attrition rate. A report from the American Academy of Sciences (Scheffler, Weisfeld, Ruby, & Estes, 1978) has called for research bearing on the problems of attracting and retaining physicians in primary care. Referring to medical education in general, Lief (1971) concluded that medical educators need to look at the affective characteristics of students.

This research was supported in part by the College of Community Health Sciences, University of Alabama, University, Alabama 35486.

Reprints and reference list may be obtained from: Kenneth J. Wade, Office of Educational Development, School of Medicine, UAB, Birmingham, AL 35294.

THE PROBLEM

This exploratory study was designed to investigate selected personality dimensions of young physicians entering family practice, as compared to the personality characteristics of consultant specialists. In order to do this, first-year family practice residents were compared with a group of first-year consultant specialty residents who were training in surgery and internal medicine. The variables used for comparison were responses to the Rokeach Values Survey, Form D, and the Rokeach Dogmatism Scale, Form E. The following hypotheses were tested:

1. Value rankings on the Rokeach Value Survey will not discriminate between family practice residents and consultant specialty residents.
2. The mean dogmatism score of family practice residents will not differ significantly from the mean dogmatism score of consultant physicians.

Studies of medical student personality characteristics present evidence that different personality types are attracted to different areas of medicine. Parker (1958) found evidence that students scoring high in authoritarianism are less person-oriented. Another study showed residents in psychiatry to have lower dogmatism scores than residents in internal medicine and surgery (Marcus, 1964.) Wasserman, Pollock & Yufit (1969) categorized medical students on a personality dimension of intimacy versus isolation. A follow-up of physicians seven years after graduate medical training showed that none of the physicians who had been rated as isolates during medical school entered a person-oriented speciality. Christies and Geis (1970) reported that medical students preferring specialties requiring more frequent interaction with people obtained higher Machiavellianism scores than students preferring low interaction specialties, such as surgery. Funkenstein (1971) concluded that before the recent emphasis on primary care training programs, people-oriented students were attracted to psychiatry.

Collins and Roessler (1975) reported results from a small sample of family practice residents (n = 11), which was compared with residents in internal medicine, surgery, OB-GYN, and pediatrics. Data had been gathered from these residents during their third and fourth years at one medical school. The family practice residents showed higher scores on need for affiliation and lower scores on aggression and materialism. Using data from the Myers-Briggs Type Indicator, Quenk and Heffron (1975) concluded that family practice residents (n = 85) obtained profiles similar to previously reported profiles of general practitioners (Myers & Davis, 1976). Family practice residents and teachers were predominatly sensing-judging types. This type individual tends to be practical and well-organized. Finally, the authors concluded that family practice was attracting a more uniform group of individuals rather than a variety of types.

The relevance of personality orientation and choice of family practice is emphasized in the literature which aims to define and distinguish family practice. Strode (1978) states that while the family physician shares medical knowledge and skills with colleagues in other specialties, his/her uniqueness is likely to be found in attitudes underscoring the primacy of the person in the patient. Blanchard (1970) and Ornstein (1977) have described the family practitioner as person-oriented. Implicit in these statements is a physician who offers patients greater autonomy and equality, a physician who uses empathic modes of observation and interaction when approaching patients. McWhinney (1975) has characterized the family physician as having the willingness to invest in relationships. Stephens (1975) has described the family physician as being able to practice medicine with a multitude of unselected patients who have unselected medical needs. Furthermore, he/she is to be able to function in a therapeutic relationship with all types of patients for an extended period of time.

Family practice, the only specialty to define its mission as exclusively primary care, places its practitioners in a client-dependent role (Robertson & Heagarty, 1975). Their patients are not likely to come from referral by another physician, but by referral from friends, family or colleagues of patients. Thus, it is more likely that the family physician will be attuned to the preferences of his/her patients rather than other physicians. The need for deriving satisfaction from patient evaluation is greater.

Finally, family practice requires a personality type which can tolerate uncertainty (Barr, 1978). One explanation for the much discussed demise of general practice has been escape from the ambiguities attending primary care. (Fox, 1957). Sharing and managing the uncertainties of the initial stages of illness call for a distinct personality type.

METHOD

Subjects

Data were collected from 135 residents in family practice, 44 residents in internal medicine, and six residents in surgery. All respondents were in their first year of residency. The family practice residents represented programs in ten states; the consultant specialty residents were training in eight different states. When considering geographic region and actual numbers of residents in the study, the majority of family practice residents were training in the southeast. A majority of the consultant specialists were training outside the southeast in states as diverse as Massachusetts and Oklahoma.

Instruments

The Dogmatism Scale, Form E consists of 40 statements, each accompanied by a six-point rating scale ranging from -3 to +3. The scale was developed by Rokeach (1960) to measure the extent to which a person's belief system is open or closed. Included in Rokeach's concept of dogmatism are the concepts of general authoritarianism and intolerance.

The Value Survey is composed of what Rokeach (1973) called instrumental and terminal values. Respondents are asked to rank order two lists of 18 values each. The first list is terminal values, the second list is instrumental values.

A cover page soliciting demographic data was attached to the Dogmatism Scale and the Value Survey. Residents were asked to give their age, sex, religion, undergraduate major, size of hometown, education and occupation of father.

Data Collection

A total of 21 family practice programs and 33 consultant specialty programs were contacted via letter or phone call to residency directors or other residency program staff. The research was briefly explained, and their cooperation was solicited. A commitment to participate meant that the correspondent would receive the survey forms, administer them to first year residents, and return them in the mail to the researcher. One consultant specialty program did not follow this data collection procedure. In this case, the program director provided names and addresses along with a memorandum of introduction for the researcher to make a direct mailing to each first year trainee. A 25% return rate was obtained from this mailing. From the residency programs initially solicited, 62% of the family practice programs returned data, 30% of the consultant specialty programs returned data. Overall, responses from family practice residency programs were much more positive toward the study than were responses from consultant specialty programs. Some consultant specialty residency directors explicitly questioned the efficacy of the study for determining the type of person entering family practice.

Data Analysis

The data from 37 variables, i.e., 36 value rankings and a dogmatism score were analyzed with a step-wise discriminant analysis (Nie, Hull, Jenkins, Steinbrenner, & Bent, 1975). The number of statistical comparisons in step-wise analysis resulted in an inflated alpha level or chance of error in identifying variables which make a significant contribution to a discriminant function. The chance of a Type I error is compounded according to the general formula $1-(1-\text{alpha})^k$. In this study, the chance of rejecting a true null hypothesis rose to $1-(1-.05)^3$ or 14%. The resulting error rate indicates that the step-wise procedure is more reliable for identifying variables which do not discriminate than for identifying variables which do.

Results

Table 2 shows the results from testing Hypothesis 1. True Friendship was the only terminal value which made a statistically significant contribution. Among the instrumental values, forgiving and helpful made a statistically significant contribution to the discriminant function. Table 1 contains median value rankings for the two groups on the three values which contributed to the discriminant function. Hypothesis 1 was rejected with $p < .05$.

The univariate F-test used to test Hypothesis 2 resulted in $F(1, 187) = 1.36$, $p = .10$. Hypothesis 2 was accepted. No significant difference between group means on the Dogmatism Scale was found. Table 3 contains summary statistics for the Dogmatism Scale.

Observation of group variances on 37 variables in the step-wise discriminant analysis revealed the consultant specialty group to have the larger variance on 31 variables. This indicated family practice residents are a more homogeneous group. The two groups were virtually identical in their distribution on six demographic variables, age sex, religion, undergraduate major, size of hometown and social class.

Discussion

These data provide evidence that young physicians choosing to train in family practice can be distinguished by their rankings of three of 36 values on the Rokeach Value Survey. The higher rankings of the two values helpful and true friendship are consistent with the ideal model of the family practitioner. Finding that consultant specialty residents rank forgiving somewhat more highly than family practice residents rank that value is not entirely consistent with the empathic model of the family practitioner.

One of the most significant findings in these data is the homogeneity of the family practice residents (*n* = 135) as compared with consultant specialty residents (*n* = 54). The fact that they respond more uniformly does not identify salient characteristics of family practice residents; however, it does infer that a distinctive group of physicians are training in this speciality. The homogeneity of family practice residents in this study confirms similar findings based on data from the Myers-Briggs Type Indicator (Quenk & Heffron, 1975).

This study should be followed by a national survey comparing family practice residents with residents in other primary care specialties, as well as tertiary care specialties. Of course, one area of investigation left untouched by this study, and one which many family practice educators must be curious about, is the impact of family practice programs on young physician values and attitudes. A pre-residency/post-residency application of the Value Survey would provide evidence regarding the efficacy of residency program components which seek to shape values and attitudes.

Continued investigation in this area has implications for resident selection and ultimately a resolution of the primary care attrition problem. With an adequate profile of the typical family physician, students considering family practice and residency directors selecting residents could maximize the outcomes of career decisions.

Table 1

Family Practice and Consultant Specialty Median Value Rankings on Three Values Which Effectively Differentiated Between the Two Groups

	Family Practice Median Ranking	Consultant Specialty Median Ranking
True Friendship	6.20	7.50
Helpful	6.18	8.50
Forgiving	9.53	7.25

Table 2

Variables Remaining in the Discriminant Function After 12 Steps

Variables	Entry F Ratio	df	Discriminant Weight
Terminal Values			
A Comfortable Life	2.66	3,185	-.48
Family Security	1.75	7,181	-.55
National Security	2.81	6,182	-.32
Social Recognition	1.18	8,180	-.20
True Friendship	6.21*	2,186	.34
Instrumental Values			
Forgiving	4.22*	4,184	-.55
Helpful	7.65**	1,187	.38
Imaginative	1.20	11,177	-.25
Intellectual	2.80	5,183	-.31
Obedient	1.10	10,178	-.28

*$p < .05$

**$P < .01$

Table 3

Summary Statistics for the Dogmatism Scale

	Mean	SD	S^2	N
Family Practice	129.21	23.09	533.15	135
Consultant Specialty	133.74	26.70	712.89	54
Total	130.50	24.19	585.16	189

PRECIS

ISSUES IN TESTING AND GRADING

MODERATOR: Jayne Middleton, Ed.D.
University of Kentucky

THE EFFECTS OF COMMERCIAL TUTORING ON MEDICAL LICENSURE EXAMINATIONS

This paper describes the results of three replications of a study of the effects of commercial tutoring upon NBME Part I scores. Implications of commercial tutoring are also discussed.

CURRICULUM BIAS: A STUDY OF A COLLEGE-CERTIFYING SYSTEM FOR STUDENTS AT TWO BASIC MEDICAL SCIENCE SCHOOLS

A model is presented that can be used to determine if a college-wide appraisal system produces unbiased assessments of student achievement for students who are enrolled in regional (multiple) schools within a given college. Individuals involved in the regionalization of health professions instruction are urged to examine their student appraisal mechanisms.

MEDICAL STUDENTS LEARNING PROCESSES AND THEIR RELATION TO TEST PERFORMANCE AND ATTITUDES TOWARD ITEM FORMATS

To increase test efficiency and to construct a test which would be viewed as fair not only by the test constructor but also by students, this study attempted to determine: (a) students' attitudes on various written item formats, (b) students' predominant test-taking pattern, (c) students' individual differences in learning processes as measured by the Inventory of Learning Processes and, (d) how the students' learning processes affected their item format preference and test performance. Results are presented and discussed in the study.

THE EFFECTS OF COMMERCIAL TUTORING ON MEDICAL LICENSURE EXAMINATIONS

Lowell K. Scott, Ed.D., C.W. Scott, M.D., Russell D. Cunningham, M.D.,
Paul A. Palmisano, M.D., Nass J. Cannon, M.D. and Sam Brown, Ed.D.
University of Alabama in Birmingham

Introduction and Statement of the Problem

In addition to the internal evaluation of medical students by faculty, some crucial decisions regarding admission, promotion, and licensure are based wholely, or in part, upon external standardized examinations, e.g., MCAT; NBME, Parts I, II, and III; FLEX; etc. In particular, the National Board of Medical Examiners (NBME) Part I represents a substantial hurdle and a crucial turning point in the academic preparation of many physicians. Not only is Part I, or its equivalent, required for licensure, but nearly half (41%) of U.S. medical schools list a passing score on the exam as a curriculum requirement.[1]

The NBME, Part I is a norm-referenced exam with the mean score set at 500 and the standard deviation set at 100. The passing score on Part I (380) is set at 1.2 standard deviations below the mean, resulting in an automatic failure rate of approximately 12%. This is obviously a highly competitive exam with serious consequences for failure; therefore, examination preparation is crucial for those students whose careers may be contingent upon a passing score.

It is no surprise, therefore, that entrepreneurs have established courses specifically designed to increase scores on the NBME, Part I and similar exams, and that significant numbers of students utilize these commercial courses as a part of their exam preparation.* With the advent of commercial test preparation, two issues of importance to medical schools, and the profession as a whole, are raised: (1) How effective are commercial test preparation courses? and, (2) Are the courses legitimate educational tools, or do they serve to circumvent the standards of the profession by providing some students an unfair advantage on the licensure exams? The wide spread impact of commercial test preparation courses and the concern over their legitimacy is emphasized by the fact that the Federal Trade Commission is currently investigating the multi-million-dollar business.[3] The question of the effectiveness of commercial test tutoring courses has not been answered by the commercial companies, who do not have access to exam results, nor by the medical schools. Lewis and Kuske[4], who have published the only significant study to date, concluded no effect due to commercial tutoring.

Research question: Does participation in a commercial test preparation course affect NBME, Part I performance?

---

*One commercial company anticipates training 25,000 students in 1979.[2]

Reprint requests: Dr. L.K. Scott, Office of Educational Development, 310 Medical Education Building, University of Alabama in Birmingham

Commercial Course

The commercial course reported in this study consisted of a series of audio tapes and supplementary written materials. The tapes were organized around items in the NBME format whereby a narrator would read a multiple choice question and give the possible responses. The student would answer the questions, and the taped narrator would explain, in detail, the logic associated with the correct response, as well as the faulty logic associated with each of the distractors. Students were expected to understand each practice item before progressing to the next, and this often necessitated supplementary reading and study. The tapes covered each of the basic sciences and students were free to select any part, or all, of the subject areas for study. In addition to basic science content, the commercial course also provided taped and written instructions in test-taking techniques.

Method

In order to prepare for the June, 1977 and June 1978 NBME, Part I exams, a total of 35 second year medical students from the University of Alabama in Birmingham (UAB) enrolled in a commercial test preparation course. These students will be referred to as the "commercial students". The remaining 271 UAB students who took the exams prepared for the Boards without benefit of the commercial course.

The 1977 (n=17) and 1978 (n=18) commercial course students were matched by basic science grade point average (BSGPA)* with equal numbers of students from the two classes, and their Part I scores were compared by non-directional t-tests for each of the years, 1977 and 1978, and for the 2 years combined. The BSGPA is an appropriate matching variable inasmuch as the Part I exam is a measure of basic science knowledge and the cumulative BSGPA is an indication of the students' basic science knowledge before the introduction of the experimental variable, method of test preparation. The correlations between UAB's BSGPA's and Part I scores were .77 and .71 in 1977 and 1978 respectively.

The commercial students were surveyed by questionnaire following each administration of the exam. Twenty-one (60%) of the commercial students responded to the surveys.

Results

Part I Scores

As illustrated in Table 1, the 1977 commercial course students scored significantly higher on Part I than did the matched students of equal BSGPA's. Although the 1978 commercial students scored an average of 24 points higher on Part I than the matched students, this difference was not significant at the .05 level (Table 2).

---

*The commercial students' mean BSGPA was approximately one standard deviation below the class mean for each year of the study.

316

Table 1 - NBME Part I Scores of 1977 Commercial and Non-Commercial Students Matched by BSGPA

	n	Part I Mean	t
Commercial	17	503	3.05*
Non-Commercial	17	434	

*p <.01

Table 2 - NBME Part I Scores of 1978 Commercial and Non-Commercial Students Matched by BSGPA

	n	Part I Mean	t	
Commercial	18	448	1.04	N.S.
Non-Commercial	18	424		

Each of the 2 years taken separately represents a one-shot sample in which the small n may have led to spurious results. However, the two years combined produces both a respectable n of 35 and the additional stability of data collected over time. The analysis of data for the combined students from 1977 and 1978 indicated that the Part I mean score of 475 for the commercial students was significantly higher than the 429 mean for the matched students (Table 3).

Table 3 - Part I Scores for 1977 and 1978 Commercial and Non-Commercial Students Matched by BSGPA

	n	Part I Mean	t	
Commercial	35	475	2.81	p <.01
Non-Commercial	35	429		

Using basic science GPA as an indicator of the students' basic science knowledge level prior to preparation for Part I, it is apparent that UAB students who prepared by the commercial method performed significantly better on Part I than students of equivalent basic science knowledge who prepared by the traditional methods.

Student Survey

Students unanimously agreed that the course helped them with the exam, and all but one noted that they would take the course again under similar circumstances. Students noted the following characteristics of the commercial course which were of great value: (1) organization of study materials; (2) content highly relevant to the Part I exam; and (3) disciplined approach to review. Also mentioned as important were, confidence gained by taking the course and experience with practice tests.

Students were in agreement that the concepts covered in the commercial course were very closely related to the concepts covered on Part I. When asked if Part I items appeared verbatim in the commercial course, 76% responded, "yes", although they noted that the percentage of verbatim items was small.

Although the course was expensive ($550), the students generally agreed that paying for the course did not significantly increase their motivation to study. The commercial students did devote more time to exam preparation than did non-commercial students (12 wks. vs. 10 wks.), however an independent study of the UAB class taking the 1978 exam found no significant correlation between exam preparation time and Part I score.

Discussion and Implications

Results of this study indicate that the examinees benefited significantly from the commercial course. Were these benefits a result of an educationally sound review, or did they result from a prior exposure to actual Part I items or "concepts"? Our survey of students suggests a combination of both.

The commercial course seems to offer an educationally and rationally sound approach to test review. The Part I exam covers approximately two years of basic science material and the typical medical school allows only two weeks of class free time to prepare for the exam.[5] The commercial company offered students a well organized, condensed, review program which was directed toward the concepts most relevant to the Part I exam. No individual medical student has the resources, time or expertise to organize as complete or relevant a review program as that provided by the company. In addition, student feedback suggests that the commercial course employed sound educational approaches to the review sessions, e.g.: instant feedback, self-paced instruction and multi-media instruction.

On the other hand, 76% of the responding commercial students reported seeing items on Part I verbatim as they appeared in the commercial course. Although NBME security is tight, and additional measures are being taken[6], it is possible for previously used NBME items to be obtained. For example, the authors understand that it is possible to obtain old exams through sources in Mexico. Another possibility for obtaining Part I items is by simply polling students after they have taken the exam. Conducted in a systematic manner, this process could lead to the reconstruction of an entire exam. Employment of this procedure over a few years could not only provide one with a good idea of the concepts likely to be covered, but could also result in the acquisition of some items which would appear on subsequent exams inasmuch as the NBME includes a small number of previously used items on each exam.[6]

The potential impact of a non-academic incursion upon the licensure and promotion of medical students is more far-reaching than may be readily apparent. The primary route to licensure for U.S. medical students is through the NBME Parts I, II and III, with Part I being, by far, the most difficult hurdle. Given the potential for a few medical students to pass Part I with the aid of commercial tutoring, the probability of incompetent students graduating remains minimal due to the stringent internal admissions and evaluation processes of U.S. and Canadian medical schools, which function independent of the external licensure process. However, if the commercial course does provide some students with an unfair advantage, the remainder of students taking the NBME exams are penalized due to the normative scoring of the exams.

The most pressing danger lies with the influx of foreign medical graduates (FMG's)* and U.S. students of foreign medical schools. Although there are obviously many outstanding medical schools outside the U.S. and Canada, there is ample evidence to suggest the substandard ability of many of the applicants from foreign medical schools who seek advanced student and residency status in U.S. medical schools and teaching hospitals.[8-10] During the period of 1970 through 1974, 67% of the Coordinated Transfer System (COTRANS) students failed Part I, compared to 12% for U.S. students.[8] The failure rate on the Federation Licensure Examination (FLEX) of Fifth Pathway students and FMG's applying in New York state (1976-77) was 29% and 39% respectively, compared to 5% failure for U.S. and Canadian graduates.[9]

Virtually all of the U.S. students from foreign medical schools and FMG's seeking admittance to U.S. medical schools and residencies, or licensure, take the Educational Commission for Foreign Medical Graduates (ECFMG), the Visa Qualifying Examination (VQE), or the FLEX. All items on the ECFMG[11], VQE[7], and FLEX[11] are drawn from the NBME's pool of previously used items, therefore, if a commercial company were able to obtain previously used NBME items, it could potentially offer all items in advance to students and graduates taking these exams. Although the NBME has taken additional security measures to safeguard its valuable pool of validated items, it would appear to be a hopeless goal given the wide usage of the exams and the high motivation in some quarters to possess the items.

The authors suggest that most NBME security problems could be eliminated by the release of all items following each exam, thereby allowing everyone access to a massive pool of items. New exams could be constructed from the item pool, with a percentage of new items added each year. This is consistent with present NBME policy, with the exception of releasing items. This method has been successfully employed on the certification exam of the American Institute of Certified Accountants as well as the Dental Aptitude Test and the Law School Admissions Tests. Releasing items would preclude the possibility of exploitation resulting from access to old Board items and would not be detrimental to any company whose test preparation course is based on sound educational principles.

---

*In 1976-77 FMG's filled 33% of the residency positions in the U.S.[7]

References

1. 1977-78 AAMC Curriculum Directory. Association of American Medical Colleges. Washington, DC
2. Ruck, D: The MCAT candidate's friend. American Medical News. Jan. 12, 1979
3. Lynch MC: Probe of coaching for scholastic tests raises issues about exams themselves. Wall Street Journal. Dec. 5, 1978
4. Lewis LA and Kuske TT: Commercial national board review programs: a case study at the Medical College of Georgia. JAMA 240: 754-755, 1978
5. Litzinger MJ and Welker WJ: National boards medical examination part I: a multipurpose tool? J Med Educ 54 Feb 1979
6. Security and credibility of national board examinations. The National Board Examiner. 24: 1-2 Jan., 1977
7. Special report on the first Visa Qualifying Examination. The National Board Examiner. 25: 2 Jan., 1978
8. Egan RL: US citizens and foreign medical schools. JAMA 232: 1028-1030, 1975
9. Stimmel B and Smith H: Career choice and performance on state licensing examinations of "fifth pathway" students. N Engl J Med 299: 227-230, 1978
10. Meskauskas JA, Benson JA, and Hopkins E: Performance of graduates of foreign medical schools on the examinations of the American Board of Internal Medicine. N Engl J Med 297: 808-810, 1977
11. Hubbard JP: Measuring Medical Education. Philadelphia, Lea and Febeger, 1978, p. 82: pp. 79-80

CURRICULUM BIAS: A STUDY OF A COLLEGE-CERTIFYING SYSTEM FOR STUDENTS AT TWO BASIC MEDICAL SCIENCE SCHOOLS

Diane L. Essex and William E. Sorlie

University of Illinois-Urbana

Within approximately the last 30 years, medical education has witnessed a number of new developments, ranging from the introduction of three year curricula (1), objectification of instruction (2,3), and regionalization of health professions education (4). Although it is often hypothesized that the innovation will contribute to improved student performance, it is not always possible to document such improvements (5,6). Instead, administrators and faculty may sometimes accept as partial evidence of program effectiveness that no significant declines in student achievement were noted following introduction of curricular change.

Assessment of student cognitive achievement is usually accomplished via paper-and-pencil instruments. Within some institutions, externally developed instruments, such as the National Board: Part I and/or II, are administered; in other institutions, internally developed certifying or comprehensive examinations are constructed by faculty. At institutions with objectives-based curricula, great care is taken in the development of internal certifying examinations; the goals are to match test items with objectives and to obtain a representative sampling of the hundreds or thousands of curriculum objectives (3,4). The task of developing an examination representative of the curriculum would appear to be especially problematic for institutions involved in the regionalization (multiple schools) of health professions education, particularly when the power of certification remains primarily with the larger parent campus or college and is not delegated to the regional campuses where a significant portion of the total student population is enrolled. In situations such as these, a "common" examination, developed cooperatively by college faculty at the parent and regional schools, is administered to all students; certification decisions are then made on the basis of student performance on the common college examination.

Despite that faculty cooperate to construct a "fair" examination, one may question whether the one examination is truly representative for all students in all schools. As Abrahamson (7) has pointed out:

> . . . curriculum is defined by the experts in such terms as "activites provided for the students by the school." This kind of definition suggests a rather static collection of learning experiences, listed or otherwise described by the institution. But in reality, curriculum is much more: it is dynamic, not static. It is the product of planning and execution; it varies with its participants, both teachers and learners; it changes in subtle ways even when it is apparently unchanging; in short, it has an existence which goes beyond the concept of a static listing or description of its formal components. Indeed, participants in a curriculum--teachers and learners--tend to endow their curriculum with a "life" of its own . . .(p. 951).

Reprints: Dr. Diane L. Essex, School of Clinical Medicine, University of Illinois, 190 Medical Sciences Building, Urbana, IL 61801.

Thus, even though faculties at the parent and regional schools may share the same general curriculum philosophy and objectives, it is clear that the implementation of the curriculum, which includes the specific instructional emphasis given to concepts, will vary from site to site. Further, it can be expected that examination items will reflect the special instructional emphasis given at each site. It may also be, then, that student performance on items might, in part, be dependent on the site at which she/he received instruction. If this kind of dependency were found to exist, the adequacy of matching items to college-held objectives to ensure the content representativeness of a "common" appraisal instrument might be questioned seriously. One implication is that a student's promotion or certification may not be a sole function of his/her knowledge of medicine, but may also be dependent on the site where she/he received instruction.

The issue of test fairness has been investigated thoroughly in relation to the selection of employees or applicants to educational programs (8). The focus of the research, though, has been on the selection of individuals. However, little research has been done to examine test fairness after individuals complete identified portions of an educational program. Thus, the present study was designed to investigate the potential bias of a commonly developed certifying examination administered to students enrolled at the parent and regional schools of a basic medical science program.

METHOD

Subjects. Data were collected from 260 first-year basic medical science students who took a basic science comprehensive in June of 1977. One hundred and eighty students were enrolled at the parent campus and 80 were enrolled at the regional campus.

Instrument. The basic science comprehensive is the primary tool used by the College of Medicine to certify first-year medical students for promotion to a clinical school. The examination, which consists of about 500 items distributed among 12 basic medical science disciplines, is constructed by the faculty of the two basic science schools. Items for the examination are either selected from a common item pool of previously used items or are newly constructed for that year's examination. In order to ensure the appropriateness of the examination as a certifying instrument for students enrolled in the two curricula, the faculties exchange their selected or newly authored questions for review; only those items that are acceptable to the faculty of both schools are included in the common examination. Items are scored dichotomously; a student's total or discipline score is simply the percentage correct either on the entire examination or on the relevant discipline items. The Kuder-Richardson reliability (KR-20) is approximately 0.92.

Procedure. Curriculum bias was defined within the context of the following question:

> Is the basic science certifying comprehensive a uniform instrument that discriminates among students from both campuses in a consistent fashion; or does the examination consist of two principal components in which students from one campus perform better on questions developed by their faculty and students at the other campus perform better on questions developed by their respective faculty?

In order to examine the potential existence of curriculum bias, the 484 examination items were reviewed to determine which ones had been newly authored for that examination. The newly-authored items (303) were then cross-classified by the discipline to which they were relevant and by the campus affiliation of the author (parent: PC, or regional: RC). Of the 303 items, 195 had been authored by faculty at the parent campus; 108 items had been authored by faculty at the regional campus. The remaining items (181) were selected from a pool of items used on previous examinations. In many cases, the site of authorship or school affiliation of the individual who selected the item could not be identified reliably. Thus, only newly-authored items were cross-classified.

Two total scores were then generated for each student; one total score reflected the student's performance on all items (195) authored by faculty at the parent campus (PC), while the second total score represented the performance on all regionally authored (RC) questions (108). From the total number of questions, five discipline/theme subscores (two per discipline/theme) were also generated: Biochemistry, Endocrinology, Host Defense, Microbiology, and Pharmacology. Only these disciplines/themes were selected for analysis because they contained a sufficient number of items to permit statistical analysis.

Analysis. Two discriminant analyses were done; in the first analysis, the overall performance (i.e., total score) of regional and parent campus students on items authored by the faculty of the respective basic medical science schools was analyzed. The discipline or theme subscores of the two student groups (PC and RC students) for the regional and parent campus-authored items were compared.

The reader who is tempted to conceptualize the design as a series of univariate t-tests should be cautious. When the dependent variables are intercorrelated, a series of univariate t-tests is subject to spurious differences and is clearly inferior to the multivariate discriminant analysis (9).

RESULTS

The results of the discriminant analysis with total scores as the dependent variables are presented first and are followed by the results of the analysis with discipline or theme subscores as the dependent variables.

Total Scores. Basic descriptive data by student group and site of authorship of item groups are given in Table 1. The data summarized in Table 1 include mean percent scores, standard deviations about the mean percent score and group differences between mean percent scores. A percent score is simply the number of questions the student answered correctly divided by the total number of questions.

Of particular interest in Table 1 is the line entitled "Total;" it can be seen that students at the parent campus earned a higher mean percent score (63.3) than students at the regional campus (57.2) on the (195) items authored by faculty. On the (108) items authored by regional faculty, regional campus students earned a mean percent total score of 66.6, while their parent campus peers earned a mean percent total score of 61.4.

TABLE 1

Mean Percent Scores, Standard Deviations and Differences Between Group Means

Site of Authorship of Disciplines or Themes	Student Group				Differences Between Group Means
	PC		RC		
	Mean	SD	Mean	SD	
Parent Campus (PC)					
Biochemistry	53.7	20.57	45.2	20.76	8.5
Endocrinology	63.4	15.93	50.4	15.51	13.0
Host Defense	56.9	19.92	47.8	19.46	9.1
Microbiology	51.6	13.11	45.3	10.32	6.3
Pharmacology	66.2	11.67	54.2	12.54	12.0
TOTAL	63.3	9.12	57.2	8.38	6.1
Regional Campus (RC)					
Biochemistry	51.0	19.63	52.7	21.12	-1.7
Endocrinology	54.1	20.76	47.9	23.08	6.2
Host Defense	68.9	21.75	62.7	21.26	6.2
Microbiology	55.5	13.60	67.3	15.00	-11.8
Pharmacology	73.6	13.27	70.5	16.66	3.1
TOTAL	61.4	8.58	66.6	10.18	-5.2

It was found that the differences in overall performance of parent and regional campus students on PC- and RC-authored items was significant (χ^2_{est} = 155.26, $\underline{df}$ = 2, $\underline{p} < .00001$). Examination of the signs associated with the group centroids (μ_{RC} = -0.6049, μ_{PC} = 1.3610) and standardized discriminant weights (v^*_{RC} = -1.65, v^*_{PC} = 1.59) revealed that parent campus students performed significantly better than regional students on PC-authored items; regional students earned significantly higher scores than did their student peers at the parent campus on regional-authored items. It was found that 84.2% of the variance in performance was associated with "true" group differences, i.e., it was possible to predict correctly in 84.2% of the cases, the school at which the student had received basic medical science instruction.

Discipline Scores. From Table 1, it can be noted that on item groups authored by parent campus faculty, students at the parent campus earned higher mean (discipline/theme) scores than their regional counterparts in every (5) discipline/theme. For example, the mean percent discipline/theme score (from Table 1) on PC-authored Host Defense items of PC students was 56.9 and 47.8 for RC students. The mean percent discipline/theme score on PC-authored Pharmacology items was 66.2 for MC students and 54.2 for RC students. Relative to performance on regionally-authored items, RC students earned higher mean discipline/theme scores that did PC students on two (Biochemistry and Microbiology) of the five disciplines/themes. For example, the mean percent discipline/theme score of regionally-located students on regionally-authored Microbiology items was 67.3 and of students at the parent campus was 55.5.

The results of the discriminant analysis with discipline/theme subscores as the dependent variables are summarized in Table 2. From this table it can be noted that performance differences among the student groups were significant (χ^2_{est} = 153.21, $\underline{df}$ = 10, $\underline{p} < .00001$). Slightly more than 82% of the variance attributable to "true" group differences was accounted for. In terms of the variables contributing most to group differences, it was found that on parent campus-authored items, the standardized discriminant weight (v_i^*) for Pharmacology had the largest absolute value -0.75 followed by Endocrinology -0.44 . Thus, on parent campus-authored items, performance in these two disciplines/themes (Pharmacology and Endocrinology) contributed most to the significant differences between the student groups.

TABLE 2

Discriminant Analysis, Discipline/Theme Subscores as Dependent Variables

Eignevalue: 0.83230

Standardized (Scaled) Discriminant Weights:

PC Biochemistry	-0.1252	RC Biochemistry	-0.0808
PC Pharmacology	-0.7487	RC Pharmacology	0.0686
PC Endocrinology	-0.4414	RC Endocrinology	0.1214
PC Microbiology	-0.1826	RC Microbiology	1.0051
PC Host Defense	-0.0690	RC Host Defense	-0.0920

Group Means on Discriminant Functions:

Parent Campus Students (n = 180)	-0.61
Regional Campus Students (n = 80)	1.36

Bartlett's V Estimation of Chi-Square: $\chi^2 = 153.21$, df = 10, $p < .000001$

Percentage of Variance Attributable to Group Differences: 82.3%

When the standardized discriminant weights applied to the regionally-authored item groups (i.e., discipline/theme groupings) are examined it can be observed that performance in Microbiology ($v^* = 1.00$) contributed most to the significant differences between regional and parent campus students on the regionally-authored discipline/theme item groupings.

CONCLUSION

Based on the analysis of student performance on items authored for a college (multi-school) administered basic medical science certifying examination, students at the parent campus performed significantly better on items written by faculty at their campus, while students at the regional campus performed significantly better on items written by faculty at the regionally-located campus. Thus, curriculum bias was found to exist. Given these findings, it appears that the test construction procedures, i.e., exchange and review of items by both faculties, are inadequate to prevent the occurrence of a curriculum dependence effect.

The occurrence of a curriculum dependence effect, could have serious implications for individual test scores. This is especially true because of the differences in percentage of questions that were generated by each school--65% by faculty at the parent campus and 35% by faculty at the regionally-located campus. Consequently, a legal issue could be raised by the student who "passes" questions from one campus, "fails" questions from the other, and "fails" the examination by achieving a total score just below the criterion score.

Given the continued move toward regionalization of medical training, increased attention should be given to the appropriate assessment of student knowledge. Students, irrespective of their campus affiliation, should have equal opportunity to demonstrate their achievement. It certainly may be, though, that a common examination, developed cooperatively by faculty at the parent and regional campuses, does not entirely serve its intended function--the unbiased assessment of student achievement. Since as Abrahamson (7) has pointed out, curriculum is the product of people, planning, execution, and resources, one might expect that curriculum dependence would exist in a common examination. Thus, participants who are involved in the regionalization of health professions instruction are urged to examine their certifying system to ensure the fair assessment of student achievement. It may be that the parent and regional schools should each separately certify their students, since faculty at the respective campus are most familiar with the curriculum, instructional emphasis, and students.

REFERENCES

1. Trzebiatowski, G. L. and Peterson, S. A Study of Faculty Attitudes Toward Ohio State's Three-Year Medical Program. _J. Med. Educ._, _54_:205-209, 1979.

2. Silber, D. L., et al. The SIU Medical Curriculum: System-Wide Objectives-Based Instruction. _J. Med. Educ._, _53_:473-479, 1978.

3. Sorlie, W. E., et al. A One-Year Program in Basic Medical Sciences. _J. Med. Educ._, _48_:371-373, 1973.

4. McLeod, M. W. The Regional Health Education Center: Illinois AHEC as an Organizational Prototype. _J. Med. Educ._, _54_:300-307, 1979.

5. Garrard, J. and Weber, R. G. Comparison of Three- and Four-Year Medical School Graduates. _J. Med. Educ._, _49_:547-552, 1974.

6. Hallock, J. A., et al. A Comparison of the Clinical Performance of Students in Three- and Four-Year Curricula. _J. Med. Educ._, _52_:658-663, 1977.

7. Abrahamson, S. Diseases of the Curriculum. _J. Med. Educ._, _53_:951-957, 1978.

8. Petersen, N. S. and Novick, M. R. An Evaluation of Some Models for Culture-Fair Selection. _J. Educ. Mgmt._, _13_:3-29, 1976.

9. Tatsuoka, M. _Multivariate Analysis: Techniques for Educational and Psychological Research_. New York: Wiley, 1971.

Medical Students' Learning Processes and Their Relation to Test Performance and Attitudes Towards Item Formats

Christine Nu Viet Vu, Ph.D.*
Southern Illinois University School of Medicine

INTRODUCTION AND PURPOSE

One of the main concerns of evaluators is to maximize test fairness and efficiency in measuring knowledge. Most decisions about item format or organization are made on the basis of what is considered to be best by the test constructor or what is recommended by test and measurement texts. Although tests are written to be administered to students, little has been documented on how students view the fairness and efficiency of item formats and item organizations. Also, little research has been done on how students' cognitive and non-cognitive individual differences affect their attitudes towards item formats and their test performance.

Test efficiency is said to be increased if the test is organized in a way which allows students to establish continuity in their response method and which facilitates their test-taking. While many authors recommend organizing test items by format (1-3), by difficulty (1-4), or by content (2,3), it has never been carefully documented whether these various arrangements facilitate test-taking for students. A second method for increasing test efficiency deals with level of thinking. It has been recognized that the assessment of students' cognitive knowledge should include assessment of both ability to recall specific facts and ability to integrate such facts and apply them in solving problems. Although some item formats may measure one type of knowledge better than others, the formats often are chosen arbitrarily.

In an attempt to construct a test which would be viewed as fair by both the test constructor and students, Wallace (5) devised a survey to poll students' opinions. The majority of students in the study indicated that written examinations should contain both multiple-choice and essay questions rather than either format alone. The same results were obtained by Huxham, et al. (6) who felt that the examination preference of students should be "...given due weight in test planning since a strong negative attitude to a given test type could be detrimental to performance" (p. 90). The purposes of the present study were to extend these findings and to determine: (1) students' views of specific written exam item formats in terms of preference, degree of difficulty, and efficiency in measuring knowledge, (2) students' predominant test-taking pattern, the knowledge of which, may allow test constructors to organize tests which are efficient for most students.

Several studies have demonstrated that students' non-cognitive individual differences, such as introversion-extroversion and neuroticism-stability are related to achievement (7,8) and item format preference (6). As little has been reported about such relationships with students' cognitive characteristics, this study investigated the relationships between students' learning processes as measured by the Inventory of Learning Processes (9) and item format preference and test performance. This present study attempts to determine: (1) the specific learning processes which differentiate among medical students and (2) whether and how these differences in learning processes affect their item format preference and overall performance in written examinations.

METHODOLOGY

The subjects were second year medical students at Southern Illinois University School of Medicine. At the end of a six month instructional period, all students were administered a Pharmacology Comprehensive Exam and the Students Questionnaire on Exams on one day, and the Inventory of Learning Processes on the next day.

---

*Reprint requests: Christine Nu Viet Vu, Ph.D., Curriculum Affairs and Educational Resources, Southern Illinois University School of Medicine, P.O. Box 3926, Springfield, Illinois 62708.

Pharmacology Comprehensive Exam (PCE) - Our medical curriculum is a system wide objectives-based instructional program; our evaluation of learning is restricted to those competencies specified in the curricular objectives. The PCE consisted of 102 items which covered most of the Pharmacology objectives taught during a six month period. All faculty provided test items in their respecitve areas and classified their own items using Bloom's Taxonomy (10). The test was divided into four levels: knowledge (55 items), comprehension (17 items), application (14 items) and synthesis (10 items).

Students' Questionnaire on Exams (SQE) - A 16 item SQE was developed to obtain students' opinions on various aspects involved in test construction. Given six different item formats (i.e. single best-answer multiple-choice, multiple true-false, true-false, situation, matching and short-answer), students were asked to rate the formats in terms of preference and difficulty on a scale from one to six. To determine the efficiency of the various item formats, students were asked to list the three formats which they believed would best assess their knowledge and understanding. The situation format (a case followed by a series of questions) was further investigated in the questionnaire to determine its advantages or disadvantages and whether it probes a higher taxonomic level of thinking. The final two items on the SQE dealt with students' test-taking pattern. One sought their preference on grouping items by format or content. The second examined their preferred sequence when faced with several tests of varying difficulty. Because it was common in our system to distribute simultaneously all tests which were to be taken within the same block of time and to allow students to proceed in the sequence they chose, it was possible to determine their preferred pattern in taking several tests.

Inventory of Learning Processes (ILP) - The ILP developed by Schmeck, Ribich and Ramanaiah (9) is a student self-report of learning processes, which consists of four scales: (1) The synthesis-analysis scale assesses "deep (e.g. semantic) information-processing habits" and organizational processes, (2) the study methods scale assesses "adherence to systematic, traditional study techniques." (3) the fact retention scale assesses "the propensity to retain detailed, factual information" and (d) the elaborative processing scale assesses "the habit of restating and reorganizing information so as to relate it to one's own experience," The four scales have acceptable levels of internal consistency, test-retest reliability, yield no significant sex differences for the average profile (10) and possess construct validity (11). As the original inventory was intended for all types of students, a few items were slightly reworded and nine items were deleted from the original questionnaire to render it appropriate for medical students in our educational system. For example, the item "I get good grades on term papers" was deleted because our students are evaluated on a pass/fail system.

RESULTS AND DISCUSSION

As all the students did not give their names on the two questionnaires, the sample size varied with each analysis, and is reported accordingly within it.

On the Pharmacology Comprehensive Exam (N=72) the scores obtained on the 102 items ranged from 59 to 92 with a mean ($\bar{X}$) of 75.99 and a standard deviation (SD) of 6.34. The students' scores obtained at the four taxonomic levels were: SD=3.785; the Comprehension level (17 items): $\bar{X}$=12.16, SD=1.427; the Application level (14 items): $\bar{X}$=11.38, SD=1.445 and at the Synthesis level (16 items): $\bar{X}$=9.29, SD=1.917.

On the "Students' Questionnaire on Exams" (N=66), the formats the students most preferred were single best-answer multiple-choice and multiple true-false. The formats they preferred least were short-answer, true-false and matching. It is interesting to note that the students were evenly divided in preference opinions on the situation format. Based also on the students' percentage ratings, it was found that the formats which were rated as most difficult by the students were short-answer, multiple true-false, and situation. The formats which they noted as least difficult were true-false, single best-answer multiple-choice and matching.

In comparing students' ratings of item formats on the preference and difficulty scales, it was found that not all preference of formats was due to level of difficulty. While the single best-answer multiple-choice might be preferred because it is rated as least difficult and the short-answer might be disliked because it is considered as most difficult, the same did not hold true for the other formats. The multiple true-false was rated as most preferred but also as most difficult, and the true-false and matching as least preferred and also as least difficult. This last finding suggested that preference of item format might depend on other factors. Students' preference of multiple true-false, for example, might be due to the fact that this format measures knowledge efficiently; students' dislike of true-false and matching might result from their perceptions of these two formats as poorer and less valid measurements of learning.

The above explanations were partially confirmed when the students were asked to list the three formats which they thought could best assess their knowledge. The most commonly chosen formats were: situation (83%), multiple true-false (67%), short-answer (65%), and single best-answer multiple-choice (53%). It was further found that the students rated the situation format as the best format in assessing knowledge because it involved more thinking than memorization. In fact, 80% of the students indicated that questions within the situation format required them to distinguish relevant from irrelevant data, and to integrate various data in order to answer the questions appropriately (88%). They reported the advantages of the situation format as helping them to reason and apply their knowledge and providing a good learning experience. On the other hand, its reported disadvantages included (a) the dependency of items within a case problem causing cumulative error, (b) the ambiguity of the questions and (c) lack of emphasis in class on applying learning in clinical cases. It is interesting to observe that with the exception of the single multiple-choice format (which has the lowest rating here), all the formats which were considered as the best measures of knowledge were also judged as difficult by the students. This observation may imply that students considered a format difficult when it requires understanding and integration of data rather than rote memorization.

Concerning students' predominant test-taking pattern, it was found that while 21% of the students preferred to have the exam items grouped by format and 20% of them preferred to have the items grouped by content, 59% of the students had no preference. In general, contrary to common textbook suggestions (1-3), students thought the arrangement of items by format or content was not of great importance. This suggests that such arrangements might not cause any great interferring effects on student test performance. When given the choice, students chose to take the exam which they considered most difficult first while they were still fresh and not tired, and to take the one they considered least difficult last. Whether these preferences are generalizable to students' approaches to test items of varying difficulty is not certain; however, these findings suggest that the traditional arrangement of items from easy to difficult may be indeed contrary to students' preference.

The results on the "Inventory of Learning Processes" (N=58) presented in Table 1 showed that students seemed to be homogeneous in the synthesis-analysis and fact retention scales and differentiated among themselves in the elaborative processing and study methods scales. The students were very similar on most of the behaviors in the synthesis-analysis and fact retention scales: they rated themselves as able to "glean organization from a unit of material" (11), to perform deep level thinking processes and to store detailed and factual information. But on the elaborative processing and study methods scales, the students' ratings indicated that they differed among themselves on abilities to encode and translate new information and on the study techniques. It was also observed that on the study methods scale, most of the students did not adopt the general pattern of regular reviews and study hours and did not attempt to gain additional information and learning on their own. This last observation could be tentatively explained by the fact that being provided with objectives indicating content behaviors to be learned and tested, most of the students did not have to adopt traditional study methods. Before these results can be generalized, additional data are still needed to determine the study habits of students who are provided with objectives versus students who are not.

In order to determine whether students' study processes have any effect on their item format preference, their ILP scores were correlated with their item format preference ratings (N=25). It was found that only the Study Methods scale was related to students' preference of single best-answer multiple-choice ($r=-.52$, $p<.05$), true-false ($r=-.39$, $p<.05$) and short-answer ($r=+.41$, $p<.05$). These results indicated that students who adopt traditional study methods such as frequent reviews, completing all assignments, and generally studying a lot tended to prefer short-answer, while the ones who did not, tended to prefer multiple-choice and true-false formats. The preference of students with traditional study methods for short-answer format could be because that format requires more studying and elaborated knowledge. On the other hand, preference of single best-answer multiple-choice and true-false formats by students who tend not to adopt traditional study methods could be because such formats do not require a lot of studying.

To assess the relationships between students' learning processes and their performance on PCE, the data (N=44) were studied from different perspectives. By correlating both the four sub-scales of ILP with the students' total performance as well as with their performance at the four levels of thinking in the Pharmacology test, no substantial correlation was found between the ILP scales and performance. Although the correlations indicated the extent and type of relationship between students' learning processes and performance, they did not provide an overall description of how the low, medium, and high scorers in written examinations differed in specific type of study processes. In order to determine whether the students' study processes profiles differed, the students' responses on each item of the ILP were plotted against their scores in Pharmacology which were divided into the top, middle and low thirds. From the obtained plot it could be interpreted that the students who tended to score higher on the total exam differed from the ones who score lower by completing all course assignments (item 22), by having the ability to memorize (items 3, 39, 40, 44), and by devising systems to recall materials (items 14, 32, 33, 46, 49, 50). When the students' scores in the Pharmacology exam were broken down into four levels of thinking: synthesis, application, comprehension and knowledge, the plot showed that the students who tended to score higher at the synthesis and application level had similar study processes. The high scorers tended to differ from the low scorers by completing all their assignments (item 22), by remembering formulas (items 40, 42) and by devising special ways to learn their materials (items 32, 33, 40, 52). In other words, students who are considered as successful in applying and synthesizing their learning in Pharmacolocy tended to follow study directions carefully and to have a good system to recall and internalize the learning content. For the comprehension and knowledge levels, as there was little variability in the students' scores on the Pharmacology exam, the profiles of the high and low scorers at these two levels were not as clear and the findings may well be spurious. It should be noted that the above descriptions of the learning processes profiles of the high versus low scorers are derived from the plot and are still preliminary. These descriptions still need to be replicated and validated and are provided here to serve as a starting point for future studies.

CONCLUSIONS

Based on the results of this study, several suggestions were made: (1) as the situation, multiple true-false, short answer, and single best-answer multiple-choice formats are considered as assesssing knowledge best, they should be considered for inclusion more often within a test, (2) the arrangement of items by format or content should not be considered of great importance, (3) the findings indicating that students preferred to take a difficult test first and an easy one last suggest the possibility that the traditional arrangement of items from easy to difficult may be indeed contrary to students' preference. In addition, as it was found that students' differences in study methods were related to their format preference, it is suggested that a test, in order to be fair, should consist of various item formats in order to allow students to have a fair chance to demonstrate their knowledge. Finally, although it was described that, on the total test and on the synthesis and application items, the higher scorers differed from the lower ones in some study processes, these last descriptions are still preliminary. Further studies are still needed to replicate and validate those descriptions.

Selected References

1. Beggs, D.L. and Lewis, E.L. Measurement and Evaluation in the Schools: Boston Houghton Nifflin Company, 1975.
2. Gronlund, N.E. Measurement and Evaluation in Teaching. New York: Mac Millan Publishing Company, Inc., 1976.
3. Sax, G. Principles of Educational Measurement and Evaluation. Belmont, California: Wadsworth Publishing Company, Inc., 1974.
4. Helmstadter, G.E. Principles of Psychological Measurement. New Jersey: Prentice Hall, Inc., 1964.
5. Wallace, W.F.M. Student Attitudes to Examinations and How they are Related to Performance. British Journal of Medical Education, 1971, 5, 273-278.
6. Huxham, G.J., Lipton A., and Cummins, R.A. Student Test Type Preference and its Relation to Personality and Achievement. Medical Education, 1976, 10, 90-96.
7. Lipton, A., Huxham, G.J., and Hamilton, D. Influence of Personality on Achievement of Medical Students. British Journal of Medical Education, 1975, 9, 215-222.
8. Huxham, G.J., Lipton A., and Hamilton, D. Reasons for Differential Performance in Multiple Choice and Essay Tests. British Journal of Medical Education, 1975, 9, 264-272.
9. Schmeck, R.R., Ribich, F.D., and Ramanaiah, N. Development of a Self-Report Inventory for Assessing Individual Differences in Learning Processes, Applied Psychological Measurement, 1977, 1, 413-431.
10. Bloom, B.S. (Ed). Taxonomy of Educational Objectives Handbook: Cognitive Domain. New York: David McKay, 1956.
11. Schmeck, R.R., Ribich, F.D. Construct Validation of the Inventory of Learning Processes, Applied Psychological Measurement, 1978, 2, 551-562.

Table 1

INVENTORY OF LEARNING PROCESSES

	True	False[a]
Synthesis Analysis		
1. I have trouble making inferences.	3.4	96.6
2. I have trouble organizing the information that I remember.	10.3	89.7
3. Even when I know that I have carefully learned the material I have trouble remembering it for an exam.	12.1	86.2
4. I find it difficult to handle questions requiring critical evaluation.	12.1	87.9
5. I do well on essay tests.	72.4	17.2
6. I often have difficulty finding the right words for expressing my ideas.	20.7	77.6
7. I have difficulty learning how to study for a course.	5.2	91.4
8. I have difficulty planning work when confronted with a complex task.	3.4	94.8
9. I often memorize material that I don't understand.	20.7	75.9
10. I have trouble seeing the difference between apparently similar ideas.	6.9	87.9
11. I can usually state the underlying message of readings.	87.9	10.3
*12. I think fast.	53.4	41.4
13. Most of my instructors lecture too fast.	8.6	87.9
14. I can usually formulate a good guess even when I don't know the answer.	86.2	12.1
15. I try to resolve conflicts between the information obtained from different sources.	81.0	17.2
16. I read critically.	81.0	17.2
Study Methods		
17. I cram for exams.	75.9	17.2
18. I have regular weekly review periods.	6.9	93.1
*19. Getting myself to begin studying is usually difficult.	39.7	58.6
*20. I review course material periodically during the term.	39.7	58.6
*21. I maintain a daily schedule of study hours.	34.5	63.8
*22. I carefully complete all course assignments.	48.3	48.3
*23. I rarely write an outline of the material I read.	63.8	34.5
*24. I spend less time studying than most of my friends.	32.8	58.6
25. I prepare a set of notes integrating the information from all sources in a course.	22.4	75.9
*26. I rarely read beyond what is assigned in class.	60.3	36.2
*27. I usually refer to several sources in order to understand a concept.	41.4	58.6
28. Toward the end of a course I prepare an overview of all material covered.	25.9	74.1
29. I increase my vocabulary by building lists of new terms.	27.6	70.7
30. I rarely use the medical dictionary.	13.8	84.5
*31. Even when I feel that I've learned the material, I continue to study it.	44.8	55.2
*32. I make simple charts and diagrams to help me remember material.	63.8	34.5
*33. I always make a special effort to get all the details.	36.2	62.1
34. I do not usually work through practice exercises and sample problems.	22.4	77.6
35. I have a regular place to study.	93.1	6.9
36. I have difficulty locating particular passages in a textbook when necessary.	1.7	96.6
*37. I would rather read a summary of an article than the original.	53.4	39.7
38. When studying for an exam, I prepare a list of probable questions and answers.	8.6	91.4
Fact Retention		
39. I do poorly on exams requiring much factual information.	5.2	94.8
*40. I am very good at learning formulas.	53.4	36.2
41. I do well on completion items.	77.6	17.2
42. I have trouble remembering definitions.	19.0	79.3
43. My memory is actually pretty poor.	8.6	84.5
*44. For exams, I memorize the material as given in the text or class notes.	39.7	55.2
Elaborative Processing		
45. New concepts (i.e. new drug, or pathology of disease...) rarely make me think of many other similar concepts.	3.4	94.8
46. While studying, I attempt to find answers to questions I have in mind.	86.2	10.3
47. I am rarely able to design procedures for solving problems.	5.2	91.4
*48. I rarely sit and think about a unit of material which I have just read.	31.0	65.5
*49. When learning a unit of material, I usually summarize it in my own words.	58.6	37.9
*50. When I study something, I devise a system for recalling it later.	51.7	44.8
51. I learn new ideas by relating them to similar ideas.	87.9	10.3
*52. I do not try to convert facts into "rules of thumb."	48.3	50.0
53. While learning new concepts, their practical applications don't usually come to my mind.	12.1	82.8

[a] These are the percentages of students' response in each category and are added here for the readers to study. The percentage of students choosing neither true or false is not listed.

*These behaviors had approximately at least one-third or at the most one-half of the students' True or False ratings.

SYMPOSIUM

NEW EXPERIMENTS TO IMPROVE PREMEDICAL AND MEDICAL EDUCATION

Chairman: John I. Sandson, M.D.

Participants: Ernest Blaustein, Ph.D.
Peter A. Stewart, Ph.D.
Arnold W. Ravin, Ph.D.
Miriam B. Rock, Ph.D.

Discussant: Mark F. Jackman, Ph.D.

NEW EXPERIMENTS TO IMPROVE PREMEDICAL AND MEDICAL EDUCATION

Beginning in 1976 a number of universities began to undertake experiments to improve premedical and medical education. At present, seven universities including Boston University, Brown University, University of Chicago, Dartmouth College, Duke University, Johns Hopkins University and University of Rochester, have received support from the Commonwealth Fund for such experiments. At some of the institutions the programs have been in place for about three years and in others the programs have been operational for about a year. The University of California (San Diego) has just been funded by the Commonwealth Fund.

Each program has its own specific goals and objectives. None of the experiments have identical objectives. All were undertaken with the belief that the present continuum of premedical and medical education can be improved. Most of the programs emphasize experiments in the middle three or four years in the usual continuum of premedical and medical education. Most have a component dealing with early selection for medical school. Some stress the importance of better counseling, especially in health-related careers. All are striving to improve the education in the humanities and social sciences in college, to minimize redundancy in science education throughout the total eight years, and to do this in such a way to encourage future physicians to develop broad interest in the liberal arts. Chart I summarizes the current status in the programs that have been operational for one year or more.

The agenda at the symposium will be to present four of the programs (Boston University, Brown University, University of Chicago, and University of Rochester) in detail. Chart I summarizing all the programs will be discussed by the Chairman after these four programs have been presented in detail. During the summary presentation the available data in the 18 years' experience at Boston University with a successful Six-Year Program will be presented as well as evaluation data that has been developed during the two-year experience with the new program at Boston University. These presentations will take no more than an hour. At least one-half hour will be available for questions from the audience.

It is hoped that this symposium will make the nature of these experiments in premedical and medical education better known so that others interested in undertaking similar experiments may share the experience currently available. Perhaps out of the seven experiments a new model for college and medical education will evolve--a model aimed at correcting certain deficiencies perceived as being important by these seven universities in 1979.

The evaluation of these experiments is complicated by difficulty in determining the effect of the college and medical education on the ultimate quality of physicians. It may well be that as long as the quality of the applicant pool remains high and as long as our evaluative tools remain poor, differences in physician quality that can be related to these experiments will be hard to elicit. However, the quality of life of the student and the ultimate quality of life of the physician may be positively affected by these experiments. Such an outcome would be persuasive to many medical educators.

COMPOSITE REPORT OF INTERFACE PROGRAMS
Based upon returned Questionnaires - April 1979

	EARLY ASSURANCE/NOTIFICATION FOR FUTURE ACCEPTANCE TO MEDICAL SCHOOL				EARLY ENTRANCE INTO MEDICAL SCHOOL		
	Boston	Chicago	Dartmouth	Duke	Hopkins	Rochester	Brown[1]
Maximum admitted each year	15	60	5	15	30	16	40
Current total	30	50	10	20	18	25	
When students enter program	end 2nd	end 2nd	early 3rd	end 2nd	end 2nd	end 2nd	1st
When admitted to medical school (if different)		after 6th	after 4th	after 4th	--	--	after 3rd
Are there new courses for "selected" students?	yes	yes (some req'd)	no	yes	yes	no	yes
Do new courses include:							
revised basic premedical science courses?	yes	no	--	yes	yes	--	yes
new preclinical science courses?	yes	yes	--	no	yes	--	no
health-related humanities and social sciences?	yes	yes	--	yes	yes	--	yes
Are new courses open to other undergraduates?	yes	yes	--	yes	yes	--	yes
Are new courses open to regular medical students?	yes	no	--	no	yes	--	yes
Are preclinical sciences taken with regular medical students?		--	--	--	yes	yes	yes

---

[1] Brown University does not have an Early Identification/Selection Program as part of its Commonwealth activities; it has a pre-existing (MED-ED) program which accepts students from high school into a 7-year Bachelor's-MD curriculum.

Ernest Blaustein, Associate Dean, College of Liberal Arts, Boston University

THE MODULAR MEDICAL INTEGRATED CURRICULUM

In 1977 Boston University undertook the planning, development, and implementation of a curriculum that was designed to integrate where desirable and feasible the curricula of the College of Liberal Arts and the School of Medicine. As part of the Commonwealth Fund's Interface Program, a new and more flexible program, designated as the Modular Medical Integrated Curriculum (MMEDIC), was developed for the premedical student. This involved the integration into the regular undergraduate program of the College certain of the basic medical sciences normally taught as part of the first-year medical school curriculum. These basic sciences were restructured into new and innovative informational blocks designed as Modules of Instruction and either partially or completely fulfilled the subject matter requirements as taught at the School of Medicine.

Following the sophomore year of liberal arts study, fifteen students are accepted each year into the Program and, at the same time, receive explicit acceptance into the School of Medicine. More importantly, however, their junior and senior undergraduate programs are modified to include modular courses which reflect subjects covered in the basic medical sciences. More than twenty modules have been developed as a result of the collaborative activity of the faculties of the College of Liberal Arts and the School of Medicine.

The evaluation and selection of candidates for this program is conducted by a Committee on Admissions made up of representatives from the faculties and administrations of the College of Liberal Arts and the School of Medicine and jointly chaired by a member from each. Data summarizing the first two classes are found in Table I.

As a consequence of this new curriculum, a new kind of student population is about to make the transition as first-year medical students at the School of Medicine in September 1979. These are students who have already completed some aspects of the medical school curriculum (Table II), having taken at least five, and in some cases as many as nine, Modules of Instruction within the sixteen courses (64 semester hour credits) normally taken during the junior and senior undergraduate years. This represents the equivalent of having completed approximately one-fourth to one-third of the first-year medical school curriculum prior to formal entry into the medical school. In many cases, these students have also had the opportunity of exploring the humanities and the social sciences in ways not normally available to the traditional premedical student, of completing undergraduate research projects, and for being eligible to receive the Bachelor of Arts degree With Distinction in a particular discipline.

Based upon the quality of the student body selected for the Program and the structure of Modules developed and taught to date, it can be stated that first-year basic medical sciences may be successfully incorporated into the junior and senior years of undergraduate liberal arts study. It is also appropriate to note that these courses may be taught either at the Undergraduate College Campus or at the Medical Center, utilize faculty from both campuses, and that the students demonstrate a level of academic achievement consistent with the standards established for students regularly enrolled in the traditional curriculum at the School of Medicine.

The Modules are generally well received by both students and faculty. Overall course organization and quality of instruction were rated highly by MMEDIC students. From the instructors' perspective, the Modules have been extremely successful in that the students were found to be of high quality and highly motivated. The relatively smaller class size (10-25 students) made the courses more enjoyable for both students and instructors.

A comparative study of the MMEDIC students' learning experiences with their peers in the six-year accelerated pathway and the regular premedical pathway was undertaken in April 1978. Preliminary data indicate that MMEDIC and six-year students described their group as a source of support and mutual intellectual exchange, while regular premedical students described their group as a source of pressure. Both groups found their freshman and sophomore Biology courses competitive. MMEDIC students reported a significantly less competitive and more cooperative learning environment in their junior year Biology courses than did regular premedical students. These data suggest that, even at this early stage in the development of the Modular program, early acceptance and curricular innovation have provided a more favorable learning environment at the undergraduate level.

TABLE I

MODULAR MEDICAL INTEGRATED CURRICULUM

SUMMARY ADMISSIONS DATA

	Sept. 1977	Sept. 1978
TOTAL APPLICATIONS	52	64
INTERVIEWS CONDUCTED	27	35
TOTAL ACCEPTANCES	15	15
TOTAL ENROLLED	15	15
MEAN GRADE POINT AVERAGE (CUMULATIVE)	3.662	3.780
MEAN GRADE POINT AVERAGE (SCIENCE)	3.638	3.801

TABLE II
MODULAR MEDICAL INTEGRATED CURRICULUM (MMEDIC)

BUSM CLASS ENTERING SEPTEMBER 1979
(E1977)

MEDICAL CURRICULUM EQUIVALENCIES

	MED MS 511 Medical Micro. & Immunology	MED MS 516 Virology	Parasitology	CLA SO 421 Medicine and The Law	CLA SO 422 Socio-Medical Sciences	MED MS 551,2,3,4 Biochemistry	MED MS 744 Experimental Des. & Biometrics	CLA SO 521 Epidemiology	CLA PS 243 Lifespan Dev't. Psychology	CLA BI 528 Physiology	CLA PS 371Z Abnormal Psych.	MED MS 500 Microscopic Anatomy	MED MS 501 Gross Anatomy Thorax, Abd., Pelvis	MEDICAL CURRICULUM EQUIVALENCIES
1	x	x		x	x		x	x	x				x	Microbiology*, Socio-Medical Sciences*, Epidemiology, Psychiatry I lectures, Anatomy I
2	x			x	x	x	x		x		x	x		Microbiology*, Socio-Medical Sciences*, Biochemistry, Biometrics, Psychiatry I & II Lectures, Histology
3	x	x		x	x		x	x	x				x	Microbiology*, Socio-Medical Sciences*, Biometrics, Epidemiology, Psychiatry I Lectures, Anatomy I
4	x	x				x	x					x		Microbiology*, Biochemistry, Biometrics, Histology
5	x	x		x	x		x	x	x				x	Microbiology*, Socio-Medical Sciences*, Biometrics, Epidemiology, Psychiatry I Lectures, Anatomy I
6	x	x		x		x	x	x				x		Microbiology*, Socio-Medical Sciences*, Biometrics, Biochemistry, Epidemiology, Histology
7	x	x			x	x	x		x					Microbiology*, Socio-Medical Sciences*, Biochemistry, Biometrics, Psychiatry I Lectures
8	x					x	x					x		Microbiology*, Biochemistry, Biometrics, Histology
9	x	x	x				x	x	x		x		x	Microbiology, Biometrics, Epidemiology, Psychiatry I & II Lectures, Anatomy I
10	x	x		x	x	x			x		x	x		Microbiology*, Socio-Medical Sciences*, Biochemistry, Psychiatry I & II Lectures, Histology
11	x	x	x	x	x	x	x					x		Microbiology, Socio-Medical Sciences*, Biochemistry, Biometrics, Histology
12	x	x		x	x		x	x	x				x	Microbiology*, Socio-Medical Sciences*, Biometrics, Epidemiology, Psychiatry I Lectures, Anatomy I
13	x	x		x	x		x	x	x				x	Microbiology*, Socio-Medical Sciences*, Biometrics, Psychiatry I Lectures, Epidemiology, Anatomy I
14	x	x		x		x	x	x	x		x	x		Microbiology*, Socio-Medical Sciences*, Biochemistry, Biometrics, Epidemiology, Psychiatry I & II Lectures, Histology
15	x	x		x			x	x	x					Microbiology*, Socio-Medical Sciences*, Biometrics, Epidemiology, Psychiatry I Lectures

*Partial Completion of Medical School Course

Peter A. Stewart, Ph.D., Professor of Medical Science, Brown University

THE MEDICAL EDUCATION PROGRAMS AT BROWN UNIVERSITY

Brown University's first medical education program lasted from 1811 to 1827 and graduated 87 physicians. The second program, from 1963 to 1975, led to the Master of Medical Science (MMS) degree. It required 6 years plus 3 summers of study, and included a 4-year undergraduate science major, a humanities minor, a Master's level research thesis, and 2 preclinical years of medical school. The goal of the program was to train medical scientists. MMS graduates took the 2 clinical years, and the M.D. degree, at other medical schools.

Brown's current medical education experiment began in 1975, with community and instructional commitments to a full 4-year M.D. degree-granting program in Medicine (PIM), along with a 7- or 8-year "continuum," the Medical Education Program (MEP), which combines the college undergraduate and PIM years. The goals of this program are to train well qualified, humane physicians, and to relieve "pre-medical pressures" by admitting students into medical school directly out of high school. MEP students constitute roughly 1/2 of each PIM class, the rest being admitted by the usual admissions committee filter from a national pool of college graduates. Continuum students may apply for confirmation of their admission to the PIM after 3 or 4 years. If after 3, their first-year PIM courses count towards their undergraduate degree requirements. A separate admissions committee screens them at this point for satisfactory academic performance. Very roughly 2/3 of the MEPs follow a "science track," majoring in biology or another science; the others follow a "liberal arts track," majoring in a humanities or social science field.

A variety of new courses and concentrations are being developed in the context of the MEP continuum, with help from an Interface Program grant from the Commonwealth Fund: A new concentration in Health and Society, with several new courses; a new concentration in Neural Sciences; a concentration in Biomedical Ethics, with several new courses; new courses in "Justice and Health Care Delivery," "Medicine and the Law," "Introduction to Human Behavior" (run by a clinical psychiatrist graduate of the MMS program); a new combined physics and mathematics course developed and taught jointly by faculty from Physics and from Applied Mathematics; a new 4-semester sequence of chemistry courses. All these concentrations and courses are open to all undergraduates, not just MEP students.

The faculty involved in the PIM, along with Biological Sciences faculty, are organized as a Division of Biology and Medicine, led by a vice-president, a Dean for Medicine and a Dean for Biology. Faculty clusters by fields of interest are organized as Sections within the Division, such as Developmental Biology, Population Biology, Physiology/Biophysics, Surgery, Medicine, etc. Divisional faculty are involved in undergraduate college teaching, graduate teaching, pre-clinical PIM courses, clinical clerkships, and continuing medical education programs, depending on their expertise and interests.

Another aspect of the continuum idea is the 8th semester program. PIM students in their last semester are encouraged to plan their clinical clerkship schedules so that it is possible for them to return to the campus and take one or more courses in any area of interest.

Students in the continuum program are generally enthusiastic about the benefits of getting out from under the "pre-med grind." They still work hard, maintain higher than required GPAs, get involved in research projects as undergraduates, and as medical students (11/12's of the participants in the 1979 Medical Student Research Day Program were continuum students). It is too early to tell (and it may never be possible!) whether they become superior physicians as a result of the continuum experience, but it is already clear that their educational experience as undergraduates is a more positive one than that of the "standard pre-med."

* * * * * * * * *

Arnold Ravin, Ph.D., Director of ASHUM Program, The University of Chicago

THE ASHUM PROGRAM

The evolution of modern society has radically altered human expectations of physical and mental well-being and has altered, too, our conception of the contribution of medicine to human health. Today the medical researcher and practitioner must knowingly join with a variety of competent specialists and professionals from other fields (ecologists, sociologists, anthropologists, social psychologists, public health scientists, medical economists, ethicists) to effect the mission to which they are committed. The skill and power generated by scientific medicine are thus to be balanced by a sense of humility and a need for interdisciplinary collaboration in the face of the social web of which medicine is a part. Moreover, the scientific basis of medicine itself continues to be so fundamentally affected by new developments in biological theory and practice, that broader and deeper training in the basic biological sciences, rather than less, are called for in the modern training of a health professional.

To deal with these changes, a new program in the Arts and Sciences Basic to Human Biology and Medicine (ASHUM) was instituted in the fall of 1977 at The University of Chicago. ASHUM is a program of integrated studies in human biology and medicine jointly offered by The College and The Pritzker School of Medicine. The Program integrates the sciences basic to medicine usually offered in the first two years of medical school with social and humanistic studies relevant to questions of health. All courses have a broad and intellectually challenging framework. Many of these courses are entirely new, having been designed for the ASHUM Program. For example, a two-quarter course on Human Morphology, analyzing the tissue and organ levels of human structure on the basis of their phylogenetic and ontogenetic origins, replaces the one-quarter course on Gross Anatomy in the traditional medical curriculum.

The Program spans four years and forms a bridge between the first two years of College and the last two years of medical school. The path leading to an eventual professional or graduate degree is not being shortened. Rather, the Program makes available, in addition to the prescribed part of the curriculum, a rich variety of electives and independent study enabling the student to make a deeply considered choice as to the specific field in which he wishes to make his contribution to health: medical practice, basic biological research, socioeconomics of health care, human ecology and public health, administration of facilities and institutions for the treatment and care of the ill and aged. The significant feature of the Program is its openness and flexibility. For the future practitioner, it means taking most of his courses in a curriculum in which scientists, clinicians, and humanists have collaborated and in the company of students of diverse skills and interests who share a common concern with health.

For the future health professional who is not a doctor of medicine, it will mean having a strong foundation in human biology and a special competence to contribute to the field of health.

Students interested in any aspect of the broad field of health care are invited to apply in the winter quarter of their sophomore year. Students accepted into the program are carefully advised and have opportunities at various stages to embark upon either a route to clinical experience and an M.D. degree or to other professional or graduate studies impinging upon the area of health. En route, the students earn a Bachelor's degree in any one of a number of disciplines and, on completion of the full four years of the Program (six years from entering College), a Master's degree in Human Biology. Some of the students in the Program are selected to complete the clinical phase (final two years) of their medical education in the University's Pritzker School of Medicine. Others may do their clinical work at other medical institutions recognizing the ASHUM Program as an appropriate base for completing the final years of medical training.

* * * * * * * * *

Miriam B. Rock, Associate Dean, College of Arts & Science, University of Rochester

THE ROCHESTER PLAN

The Rochester Plan has focused upon improving undergraduate premedical and health professions curricula and providing opportunities for study which satisfy traditional pre-professional requirments while adding breadth and flexibility to student programs. It is promoting undergraduate inquiry into medically related fields and offering substantive alternatives within the health professions. A large interdisciplinary group of faculty members and administrators have stimulated and devised new interdepartmental course offerings and degree programs, supported the revision of introductory course sequences in the biomedical sciences, and expanded independent research opportunities. New lines of communication and cooperation are being established among existing departments and colleges. The resources of the entire University are being used. This consolidation of effort by undergraduate and medical school faculty, with cooperative use of facilities, has increased the integration of health science and liberal arts education.

The University's counseling facilities and services have been strenghtened. Health professions advisers in an expanded Career Services and Placement Center work closely with academic advisers, psychological counselors, and faculty from across the colleges helping students shape their vocational goals by directing them individually toward full use of University resources.

The Rochester Plan provides early admission to the University of Rochester School of Medicine for up to 16 students at the end of their sophomore year. After completing two years of liberal arts and science education, these students design individual four-year programs integrating undergraduate and preclinical medical education, followed by two years of clinical study. Although the Early Selection student spends the usual eight years required for the B.A. and M.D. degrees, the integrated four-year "middle" period allows for broad and in-depth study not possible in the conventional pre-medical/medical school sequence.

New academic opportunities supported by personalized advising have increased the number of individually structured interdisciplinary degree programs. Through the Center for Special Degree Programs, undergraduates have initiated and completed concentrations concerned with health and society issues including Medical Ethics, Death and Dying, the Politics and Economics of Health Care, as well as intercollegiately structured programs in neuroscience and biochemistry. In addition, the Department of Microbiology in the School of Medicine now offers a Bachelor of Science degree program for undergraduates in the College of Arts and Science. Students selecting these programs are not only pre-meds and Early Selectees; they are preparing for graduate study and careers in a variety of health care and science areas.

The Rochester Plan is being carefully monitored as it goes through its various phases. Traditional structures of the University--departments, colleges, department chairmen and deans, and their administrative committees--have continued to serve as the primary mechanisms for implementation of the Plan's activities.

However, the original focus of the Plan has broadened; areas outside the health sciences have been influenced and changes effected. The idea of the Rochester Plan has moved beyond the original emphasis on the health professions.

It is difficult to use only the existing structures of the University to deal with concerns which might involve ten or more faculty members from four or five departments in two or three colleges and a potential student body of undergraduate, graduate and professional students. To meet this need, a new configuration known as "clusters" has been devised. Clusters are mechanisms of interaction that promote and stimulate scholarly activity among various departments across the colleges of the University.

During the past year genetics, bioethics, aging, human and mammalian biology, health services, and environmental sciences have been identified as possible clusters. Cluster interests include collaborative research, teaching, curriculum development, faculty recruiting and development, recruiting students (undergraduate, graduate, professional and/or post-doctorate), sharing of facilities and equipment, and external funding. Yet, a cluster neither replaces nor supercedes traditional administrative structures. Graduate students recruited by a University cluster have to be admitted to a department. A Cluster may identify faculty to be recruited, and serve as an attraction in the process, but the appropriate department must approve the appointment, hire the person, and take responsibility in the review for tenure. The Clusters' primary responsibility is to catalyze and facilitate these activities.

The Genetics Cluster is one example. It was formed from small and isolated groups of faculty, a total of thirty-four in five departments in two colleges. Formerly, they recruited their own students, housed their own equipment, pursued related, even overlapping research interests, and sought out their own funds for research. The Genetics Cluster now brings together existing resources, enhances the development of genetics courses and new programs, and allows for collaborative investigation at the forefront of genetics research, thereby benefitting the faculty involved, their departments, and their students, and the University as a whole.

Each cluster is not intended to be a permanent structure of the University. The Rochester Plan Policy and Planning Committee, composed of a faculty representative from each of four areas, the physical and engineering sciences, the biological sciences, the social sciences, and the arts and humanities, plus an associate dean from the College of Arts and Science and from the School of Medicine, stimulates, guides, monitors and reviews the activities and membership of each Cluster. The administrative work of a Cluster is handled by its own steering committee. The Policy and Planning Committee reports directly to the Provost, who as chief academic officer of the University oversees all intercollegiate activities.

The Interface activities for improving education for the health professions continue to be an important component of the Rochester Plan. The success of the Interface experiment is providing the impetus for developing much broader and deeper ties of communication and collaboration within the entire University community.

SYMPOSIUM

RESEARCH IN RESIDENCY TRAINING FOR PRIMARY CARE

Chairman:	Edwin F. Rosinski, Ed.D.
Participants:	Edwin F. Rosinski, Ed.D. Janet T. Pozen, Ph.D. Evelyn R. Dienst, Ph.D. Leighton E. Cluff, M.D.

347

OBJECTIVE OF THE SYMPOSIUM

The major objective of this symposium on Research in Residency Training for Primary Care is to explore the types of research models which can be used to measure the outcomes of residency training programs. At a time when national health policy has placed a premium on primary care, university medical centers are attempting to accommodate this policy by developing a variety of approaches in the training of primary care residents.

Perhaps no medical education endeavor since the curricular reforms of the late 1950's and early 1960's has received as much attention as primary care. This symposium focuses on this endeavor by exploring the multiple avenues open to researchers in medical education in examining the components of residency training. Furthermore, the results of the analysis of the training programs have implications for the development or modification of future policies related to primary care. This symposium deals with one of the most timely medical education issues of the day.

Edwin F. Rosinski
School of Medicine
University of California
San Francisco, California

INTRODUCTION

The Robert Wood Johnson Foundation provided the initial stimulus for the development of residency programs in internal medicine and pediatrics directed toward primary care. The Foundation realized that the goal of improving the assessibility to primary care could be achieved in several ways, but it focused its early attention to the development of primary care residencies in internal medicine and pediatrics as one of several options.

The initial nine programs it supported included Boston University, Dartmouth, University of Rochester, Harvard, George Washington University, University of Missouri-Kansas City, University of Florida, and the University of California at Los Angeles and at San Francisco.

As a condition of awarding each grant, the Foundation stipulated that an "evaluation or research" component be included in each grant proposal. The Foundation did not, however, prescribe what should be studied or how the study should be carried out. Each institution was given a free hand to design its own study with the expectation that the study findings would be used to improve the quality of the residency program and the services it provided, and that the findings would add to a better understanding of the educational process in primary care residencies.

Therefore, each one of the nine programs appointed an individual to head the evaluation or research effort, and for the duration of the grant support studies on residency training continued.

Aware that each of the nine programs might undertake a different research tack - indeed the Foundation encouraged diversity - the Foundation recognized the need to undertake a study of nine programs which would provide uniform data. The data from such a study might be helpful in making policy decisions, but more than anything else the Foundation recognized the need for objective data on residency training for primary care.

The Foundation sought out the Office of Medical Education of the School of Medicine of the University of California, San Francisco, to undertake the study. The Foundation gave the Office a "free hand" on how to design and carry out the study, and requested that for its purposes three general issues be examined. These included: (1) A description and analysis of Foundation supported primary care residencies in internal medicine and pediatrics. (2) An analysis of cost and sources of support for such programs. (3) A documentation of the process which these programs were being "institutionalized," that is, becoming fully operational and institutionally supported and accepted educational programs.

To address these three major issues, as well as others the investigators themselves proposed, a study was designed and twenty months were assigned to carry it out.

What will now follow is a brief description of the research approach undertaken by two of the nine programs. Boston University had two separate residency tracks; one in pediatrics primary care, and one in internal medicine primary care. The Unversity of California, San Francisco, likewise, had a separate residency track in pediatrics and one in internal medicine. Although both of these programs had similar tracks, the emphasis of the research was decidedly different as will be seen.

The description of the research efforts of these two programs will be followed by a description of the "national" study of all nine of the residencies. Finally the entire research effort, as well as research on primary care in general, will be put into historical and philosophical perspective.

Janet T. Pozen, Ph.D.
Department of Socio-Medical Sciences and Community Medicine
School of Medicine
Boston University
Boston, Massachusetts

EVALUATING THE BOSTON CITY HOSPITAL EXPERIENCE IN PRIMARY CARE RESIDENCY TRAINING: A FIVE YEAR ASSESSMENT

Beginning in 1974, the Departments of Medicine and Pediatrics at Boston City Hospital developed primary care pathways within their residency training programs. Initially 11 percent (4/36) of the internal medicine resident positions and 31 percent (4/13) of the pediatrics resident positions were allocated to the primary care pathway. By 1978 these proportions had increased to 29 percent (9/33) in internal medicine and 50 percent (6/12) in pediatrics. The departments maintained a single internship match to enable application of equal standards to selection of both types of resident and to minimize polarization which can result from separate selection processes. The primary care and traditional pathways differ in their emphasis on training in general versus subspecialty care and in ambulatory versus inpatient care. Both the quantity and content of the ambulatory care experience are expanded for primary care trainees. Primary care clinic sessions and their accompanying preclinic conferences are precepted by an interdisciplinary faculty that includes general internists and pediatricians, primary care nurse-clinicians, social workers, psychiatrists, and social scientists. Teaching is focused on the assessment and longitudinal management of patients' medical and psychosocial problems. Required and elective time are provided for ambulatory training in other specialties, principally through a medicine - pediatrics cross-over.

Our strategy for evaluating the Primary Care Residency Training Program has generally focused on a longitudinal study of residents in the primary care pathway compared with residents in the traditional pathway as they enter and move through their training and embark on their professional careers within Medicine and Pediatrics. Through the use of self administered questionnaires and standardized test we have

(1) established the comparability of their entry characteristics,
(2) isolated similarities and differences in salient aspects of their ambulatory and inpatient care training experiences (e.g., goals for training, development of abilities and skills and qulaity of learning interactions), and
(3) compared stability and change in the development of their professional career plans during and subsequent to training.

As of spring 1979, five cohorts had been entered into the study. The first cohort had been out of the training program for two years, while the fifth cohort had just completed its internship (PG4-1) year.

Longitudinal Study of Residents

Cohort	Beginning Internship	End Internship	End JAR	End SAR	1 yr Post Training	2 yrs Post Training
1	1974	1975	1976	1977	1978	1979
2	1975	1976	1977	1978	1979	1980
3	1976	1977	1978	1979	1980	1981
4	1977	1978	1979	1980	1981	1982
5	1978	1979	1980	1981	1982	1983

Through this comparative study design we have begun to assess whether an increasingly common factor of structural modification in residency training - the primary care pathway - achieves its goal of addressing the problem of specialty maldistribution by influencing the residency training experience and resulting career choice.

Our second evaluative strategy consists of assessing the implementation of the primary care "curriculum" components. Using primary care residents' reports in the yearly survey and in specially focused studies we establish whether the curriculum components (e.g. cross-over rotations, preclinic conferences, clinic sessions, joint practice with primary care graduate - nurse trainees) as they are structured, convey a learning experience in areas identified as program objectives. (e.g., assessment and management of psychosocial and physical aspects of patients' problems).

A third evaluative strategy has been more difficult to implement. In assessing the impact of the two types of training on the content and quality of patient care provided by residents, we are limited by design possibilities and by the availability of criteria for wholistic care of specific problems. In one study we compare the quality of primary care and traditional pathway pediatric residents provide to otitis media patients in the same setting. Explicit criteria were developed for history, physical exam, medications, and followup.

Evelyn R. Dienst, Ph.D.
Division of Ambulatory and Community Medicine
University of California
San Francisco, California

EVALUATION MODEL OF THE PRIMARY CARE RESIDENCY PROGRAM

The Primary Care Residency Training Program at the University of California (UCSF), San Francisco is administered through the Division of Ambulatory and Community Medicine (ACM). The ACM Division is also responsible for administering the outpatient medical clinics at the University's hospital where residents in the traditional medicine residency program spend their half-day-a-week in the outpatient setting. The Division also coordinates the UCSF affiliated Family Practice Residency programs in Fresno, Santa Rosa, San Francisco General Hospital and recently, in Salinas.

The Primary Care Program involves residents in pediatrics and medicine who have selected for the primary care program by procedures that are distinct and separate form those used to select residents for the traditional medicine and pediatric residency programs (although some of the criteria for selection inevitably would show some overlap).

The evaluation component of the program was developed in conjunction with the training aspects and was designed to be integral to the program and its development. Both formative and summative evaluation purposes were built into the design of the evaluation project.

Comprehensive program evaluation is essential to the development of innovative programs and initially provides important information to the larger interested community on the progress of its development. Nevertheless, to accomplish both formative and summative purposes a wide range of evaluation techniques may need to be employed and may compete for resources, expertise, visibility and relevance within the developing program. Because evaluation is distinct from research, the resolution of these issues is value-laden and has implications for the collection of data and for the development of the program being studies.

The evaluation component of the primary care program was conceived as a complex process investigating the program from various points of view in order to evaluate the program's multiple objectives or tasks. The two major areas of focus have been the evaluation of "the program" and the evaluation of the "house officers."

In the evaluation of the primary care house officers, we have investigated a variety of entering characteristics of residents in the primary care program and compared primary care residents to residents in the UCSF affiliated programs in traditional medicine, traditional pediatrics and family practice. In this way we sought to identify the personality, attitude and training goal orientations of residents who entered a program with a somewhat unique primary care focus. In addition to assessing the characteristics of residents upon entrance, we have also developed a number of procedures to assess their competence as they progress through the program. We have used faculty ratings, chart reviews and patient interviews as well as observation.

The program-centered component of the evaluation focuses on the training program itself and on evaluating the primary care clinic as a health care delivery system. To evaluate the training program, residents and faculty complete questionnaires and are interviewed on a yearly basis. The results are summarized and made available to faculty and residents as part of our efforts at program development. Residents complete a daily log of their primary care clinic activities and encounters with patients. These are tallied to assess the adequacy of the clinic population to meet primary care training objectives. A follow-up study of program graduates will provide an assessment of the career direction and activities of program graduates in comparison to those who are in family practice on the traditional medicine and pediatric programs.

From the perspective of the primary care clinic as a health care delivery system, chart reviews and patient interviews assist in the assessment of the quality of care provided. A comparison of the Primary Care Clinic population with the population of the two university general medical clinics, enables an assessment of the degree of "primary care" orientation of the clinic and the nature of its patient population.

Preliminary results from the evaluation study will be highlighted with regard to the following:

Personality
Attitudes
Knowledge
Professional self perceptions
Goals for training
Evaluations of Residents Clinical Performance
Clinic Work Environment
Patient and Visit Characteristics
Patient Satisfaction

These findings indicate that residents in primary care are distinct in their committment to primary care training and have distinct personality and attitude characteristics which appear to make them expecially well suited to their goal of becoming primary care physicians.

This general orientation is not, however, without some conflicts in professional identity and these, as well as expressed concerns for certain needed improvements in the training program have been important considerations in the program's development. In general, the training program and the primary care clinic work setting received positive ratings from residents and was percieved to provide a distinct training and clinical experience. Residents as a group, are highly regarded by facluty who evaluate their performance within the primary care clinic, in subspecialty clinics and on the wards.

Edwin F. Rosinski
School of Medicine
University of California
San Francisco,California

A NATIONAL STUDY OF RESIDENCY TRAINING FOR PRIMARY CARE

As indicated in the Introduction, the Robert Wood Johnson Foundation (RWJF) sought answers to three questions in their national study. The investigators enlarged on these three questions to a point where a detailed study was designed which required the use of several techniques to obtain the data.

The study addressed itself to the following topics or issues: (1) Some of the problems encountered in developing primary care residencies and how the problems were overcome. (2) The objectives of the programs and how, and by whom, were the objectives developed. (3) The clinical affiliations used as training sites and the kinds of other training programs included at those clinical sites. (4) The types of staff used in the programs including their clinical and psychosocial orientation. (5) The primary care residents in training, considering some twenty five different variables or features regarding them. (6) The patient population presented at the clinical affiliations including social class backgrounds. (7) The content of the training programs including time spent in out-patient and in-patient activities. (8) The specific educational and patient care activities in which the residents engaged. (9) The evaluation approaches used in each program. (10) Funding sources and other funding information. (11) The process by which the programs became "institutionalized."

To examine these topics or issues two approaches were used; (1) mailed questionnaires, and (2) site visits to all programs for the purpose of having interviews with the key participants. Two separate detailed questionnaires to all nine program directors were mailed and returned. One hundred and sixty questionnaires were mailed to primary care residents and 159 were returned. Nine extensive funding questionnaires were mailed to program directors or fiscal officers and eight were returned. A total of 213 individuals, all associated with the programs, were interviewed using a structured interview schedule. Those interviewed included faculty from all academic ranks, other residents who worked with primary care residents, nurses, nurse practitioners, physician assistants, and the like. Clinic directors were also interviewed as well as the dean or vice president for health (medical) affairs. The interviews with the deans lasted at least one hour and the purpose of the interviews was to identify the forces that supported or impeded the development of the residency programs as well as the forces that were instrumental in the "institutionalization" of the programs.

The study report begins with a brief consideration of some of the issues surrounding primary care, including a summary of some of the more popular current definitions of the term. An entire chapter is devoted to a description of the nine programs and only in that chapter are the programs identified by name. The bulk of the study is devoted to the eleven topics or issues cited above.

With a wealth of anecdotal data collected from responses to such questions as "What incidents or events caused residents to select primary care?" "How residents feel about the training program?" and "Would you reapply to the program?", an extended appendix is included containing the verbatim responses.

Because of the paucity of data on individuals who select residency training in primary care, a considerable portion of the report is devoted to the resident population. This includes, among other things, a comparison of them with residents in straight internal medicine, the medical schools they attended, what curricular tracks or options they followed, their social/economic background, how they were perceived as clinicians and being knowledgeable in science of medicine, and in what other ways were they different from residents in straight internal medicine and straight pediatrics.

The study concludes with a chapter that is devoted to making some general observations regarding the eleven topics or issues. In addition, a brief section in this chapter attempts to conclude as to what are the distinctive features of the nine programs.

Wherever possible, objective, analytical data are presented and such data can be examined in the 42 tables and three figures which are included in the report.

Leighton E. Cluff, M.D.
Vice President
The Robert Wood Johnson Foundation
Princeton, New Jersey 08540

SUMMARY OF COMMENTS ON PRIMARY CARE TRAINING

Internists (generalist or specialist) provide a great deal of primary care, but residency training prior to the past few years was overbalanced in favor of hospital care for complex medical problems. Programs in primary care training in internal medicine have begun to correct this imbalance. Internists largely care for elderly patients who account for a major proportion of patient hospitalizations. The proper balance between primary care and hospital care in the training of internists, therefore, is only now being determined. Chronic disease affects most patients cared for by internists, but training in internal medicine has not yet dealt successfully with the long term management of patients with these problems. Much of the primary care training of residents in internal medicine is being provided in the teaching hospital outpatient department. The adequacy of this setting as currently in place in many institutions is questionable.

Practicing pediatricians are providing less hospital care of children than was the case in the past. The problems of children in the practice setting are not well represented in the usual hospital training of pediatric residents. For this reason there probably has been an even greater disparity between the training of pediatric residents in primary and hospital care than has been the case in internal medicine. This imbalance in pediatric training is also now being corrected.

The role of internists in care of children, the role of pediatricians in care of adults, and the training each should receive to prepare them for their prospective roles is unresolved.

It is clear that the evaluation of primary care training in internal medicine has attracted highly qualified residents, but the impact of this training upon the future function of internists is uncertain.

SYMPOSIUM

UNITED STATES CITIZEN FOREIGN-TRAINED MEDICAL STUDENTS ACCEPTED WITH ADVANCED STANDING INTO AMERICAN MEDICAL SCHOOLS: BACKGROUND, REENTRY AND PROGNOSIS

Organizer:	Paula L. Stillman, M.D.
Chairman:	Ethel Weinberg, M.D.
Participants:	Lester M. Geller, Ph.D. Paula L. Stillman, M.D. Jane S. Ruggill, M.S.W. Paul J. Rutala, M.D. Sarah M. Dinham, Ph.D. Darrell L. Sabers, Ph.D. Barry Stimmel, M.D.

SYMPOSIUM

UNITED STATES FOREIGN-TRAINED MEDICAL STUDENTS ACCEPTED WITH ADVANCED STANDING INTO AMERICAN MEDICAL SCHOOLS: BACKGROUND, REENTRY AND PROGNOSIS

Organizer:	Paula L. Stillman, M.D. University of Arizona College of Medicine
Chairman:	Ethel Weinberg, M.D. National Board of Medical Examiners

Objectives:

The goal of this session is to present information about U.S. citizen-foreign-trained medical students (USFMS) accepted with advanced standing into American medical schools.

In order to plan sound educational programs for transfer students from foreign medical schools, the curricula and learning environments of these schools must be analyzed. Dr. Geller will present the strengths and weaknesses of curricula at foreign schools and the self-study techniques used by the USFMS to compensate for deficits in their instruction.

Dr. Stillman will present objective data regarding the patient assessment skills of the USFMS upon admission to the University of Arizona College of Medicine. Major deficiencies were noted and were diminished by a brief, intensive, competency-based instructional program. Teaching techniques and evaluation strategies will be described.

Dr. Stimmel will compare the performance of USFMS accepted with advanced standing into the 3rd year of an American medical school with students enrolled in American schools (USMS). In addition, he will compare USFMS and USMS with Fifth Pathway students.

Dr. Weinberg will moderate and critique the presentations.

CURRICULUM AND LEARNING ENVIRONMENT IN FOREIGN MEDICAL SCHOOLS

Lester M. Geller, Ph.D.

Department of Pathology
College of Physicians and Surgeons of Columbia University

Despite the appearance of a comprehensive curriculum, it is important to distinguish between the offerings of a foreign medical school and the quality, quantity and type of training received by its students. Quite likely, these will be perceived and assessed differently by nationals and United States Foreign Medical Students (USFMSs). It is also helpful to differentiate between deficiencies in the USFMSs' medical education (i.e., in the basic science and clinical training made available to them), and those manifested by the students at a given time. Considerable compensation for the former frequently takes place, so that, for example, were one to judge deficiencies in basic science training by performance on Part I of the National Board, in selected groups of USFMSs, it might (could) be concluded that they hardly exist.

The impressions to be presented of the learning environment in foreign medical schools, their curricula, deficiencies and the compensatory activities of USFMSs, have been gathered from interviews with, and questionnaires completed by, approximately 250 students, and by study of school transcripts and bulletins. These data were collected during five Basic Medical Sciences Review Courses for USFMSs. Since 77% of the students came from Mexican, and 16% from Italian schools, remarks will be restricted to them. However, with some differences, they apply to virtually all schools attended by our students.

American students are accustomed to having educational opportunities scheduled and officially available. In the foreign schools, they must vigorously seek out chances to secure the training they believe necessary. How well they do this, in the host country or through externships in the U.S., and how zealously they study on their own, help to account for the differences in preparation found among USFMSs.

Aside from a weak research orientation, the major problem in the Mexican schools attended by our students lies in their faculties. Many instructors are said to be less than adequately informed in their subject, limited in experience, uncaring about students or teaching, authoritarian, and unwilling to admit to error. Coverage is sometimes superficial, and the inclusion of minutiae common. Students are occasionally suspicious of their instructors' accuracy, and always, they are anxious about preparation for the Boards and transfer to a U.S. school. Feeling, excitement and a positive attitude toward the learning environment, are lacking.

Faculty do not help to distinguish the trivial from the important, and there is a general failure to relate basic science and clinical areas. Preclinical subjects are taught and learned in a discipline-bound fashion, with emphasis upon memorization rather than understanding in fundamental biological terms. There is dismay about the inadequacy of the teaching and the meaninglessness of poorly supervised laboratory exercises. Frequently, too, equipment and facilities are seen as poor.

Deficiencies in clinical training and programs to compensate for them, developed by the schools or undertaken by the students on their own, will be discussed orally.

The situation in the Italian schools is similar in several respects to that already described, but different in others. There is a long tradition of medical training in Italy. Moreover, the state-supported schools are hardly interested in attracting students from the U.S. Faculty are viewed as competent, but, occasionally, outdated. Laboratory exercises are often thought inconsequential. Certain schools are greatly overcrowded, and individual attention is meager. Even more than in the Mexican schools, the American student in Italy is very much on his own. Patient contact and scheduled, supervised clinical experience that allows for "hands on" activity, are virtually absent. The lack of personal involvement, of doing, permeates the curriculum. Students learn about medicine from books, lectures and demonstrations, but have little opportunity to test their knowledge or skills in real situations. Emphasis is placed on "theory" as opposed to practice. Much detail, essential for passing the oral examinations, is presented in the basic sciences, with little attention paid to its clinical significance.

Overall, the deficiencies reported, and the compensatory self-study undertaken, result in primarily book-learned students. They do not always know what else there is to know, and there is difficulty in separating the important from the unimportant. Clarification of unresolved questions is hard to come by. Alternative interpretations, generally secured from interaction with knowledgeable faculty, are lacking; the textbook is their Bible, and student beliefs are the author's. Guidelines for study and for direction of effort, and a broad, integrating, conceptual orientation toward the study of medicine are missing.

Student attitudes toward their education, the positive aspects of their experiences, and elaboration of points mentioned, will be included in the oral presentation.

A COMPETENCY BASED PROGRAM TO REMEDIATE DEFICIENCIES OF COTRANS STUDENTS

Paula L. Stillman, M.D., Jane S. Ruggill, M.S.W., Paul J. Rutala, M.D.,
Sarah M. Dinham, Ph.D., Darrell Sabers, Ph.D.
University of Arizona College of Medicine

In 1978, the University of Arizona College of Medicine (UACM) changed from a three to a four-year curriculum. As a result, the College was able to accept an entire class of 41 transfer students with advanced standing. The class was composed of 30 United States citizen foreign-trained students (USFMS) and 11 United States-trained transfer students (USMS), all of whom had completed a minimum of two years of basic science courses at their previous medical schools and had achieved acceptable scores on Part I of the National Board of Medical Examiners Test (NBME I). The transfer students entered in September, 1978, and constitute the graduating class of 1980.

Based on past experiences, it was expected that students transferring into the UACM in September, 1978, from foreign medical schools might have difficulty during their initial clinical clerkships. This expectation provided the impetus for the development of an intensive six-week training and evaluation program which each student was required to pass prior to entering clinical clerkships. Criterion levels of knowledge and clinical competencies were identical to those established for students completing their pre-clinical years at the UACM. It was assumed that if the USFMS were able to meet the UACM criteria, they would be more successful in clerkships than previous students tranferring from foreign medical schools.

The training program consisted of an initial, mid-program and final evaluation and an on-going instructional program. The initial assessment of each student's clinical skills served to screen for major deficiencies. The mid-program evaluation served as a diagnostic evaluation of performance. The final evaluation determined how well each student was able to utilize skills to assess real patients. Group and individual instruction based on documented areas of weakness continued throughout the program. Because many of the instructional components were competency-based, instruction and evaluation were integrated throughout the training program.

The unique composition of the class of 1980 provided the faculty with the opportunity to examine the following questions: (1) Were USFMS, transferring into American medical schools with advanced standing, deficient in patient assessment skills? (2) Were the skills of transferring USFMS different from those of transferring USMS? (3) If, indeed, differences in clinical performance were present, could they be eliminated by a brief but intensive competency-based training program? At each stage of the program the USFMS were compared to the USMS. In addition, wherever possible, the entire class of 41 transfer students (Class of 1980) were compared to the 89 students in the UACM sophomore class (Class of 1981) or to previous classes at UACM when data on the Class of 1981 were not available.

PRE-CLINICAL DATA

Baseline data obtained for each student prior to admission to medical school included MCAT total and subtest scores and the NBME I. There were no significant differences between USFMS and USMS on either test. The class of 1980 as a whole scored lower (p=.01) than the class of 1981 on the MCAT. On

NBME I the class of 1980 was compared to that of the UACM classes of 1975-79. The class of 1980 performed significantly better (p=.05) on the physiology, biochemistry, and pharmacology subtests; the only subtest on which the UACM five-year average was significantly higher (p=.05) was behavioral science.

INITIAL ASSESSMENT OF CLINICAL SKILLS

Although all transfer students had passed a course in physical examination and medical interviewing at their former medical schools, it was deemed necessary to obtain objective and standardized baseline data in these skill areas during the first week of the training program. The USMS obtained significantly higher scores (p=.01) than the USFMS on the initial assessment of the complete physical examination and medical interview.

CURRICULUM: REMEDIAL INSTRUCTION BASED ON DOCUMENTED AREAS OF WEAKNESS AND INSTRUCTION IN ADVANCED SKILLS

The instructional program was based on predicted group strengths and weaknesses as well as on individual deficiencies identified during the initial assessment phase. The educational program included: (1) instruction in physical examination techniques; (2) lectures and small group practice sessions in medical interviewing; (3) sessions with trained patient instructors (PIs) who taught and evaluated each student's performance of the female genital and breast examination; (4) a half-day patient care skills workshop; (5) clinical correlation lectures; and (6) a preceptorship program.

MID-PROGRAM EVALUATION OF SKILLS

During the mid-program evaluation of clinical skills, each student (1) performed a complete physical examination on an asymptomatic PI, (2) elicited a complete history from a PI programmed with information about a specific medical complaint, and (3) performed a screening neurological examination on an asymptomatic PI. USMS outscored USFMS on the physical examination (p=.03), but there were no differences between the two groups on the content of the medical history elicited from the PIs, the process of the medical interview, or the neurological screening examination. The transfer students scored higher than the Class of 1981 on the process of the medical interview (p=.025), the content of the medical interview (p=.001) and the screening neurological examination (p=.02); there were no differences on the physical examination of the asymptomatic PI.

FINAL EVALUATION

The final evaluation of each student's clinical performance included four components: examination of symptomatic patient instructors, preceptor evaluation, clinical correlation examination, and complete assessment of an unknown patient. Successful completion of each aspect of the final evaluation served as a prerequisite for entrance into the clinical clerkships which constitute the junior year of the UACM curriculum.

There were no differences between USFMS and USMS on any of the PI evaluations. The transfer students scored higher than the UACM students (Class of 1981) on the pulmonary performance (p=.04), pulmonary content (p=.04), and musculoskeletal performance (p=.01); they did not differ on cardiovascular performance and content nor on musculoskeletal content.

At the completion of the preceptor program each preceptor assessed student skills in physical diagnosis and problem solving. These 41 evaluations and those collected on the 89 students in the Class of 1981 were read and rated by an independent evaluator. There was no significant difference between USFMS and USMS but the Class of 1981 was rated significantly higher (p=.01) than the Class of 1980.

The one aspect of the final evaluation in which the USFMS and USMS differed was the short answer multiple choice examination covering information presented during the clinical correlation lectures. In this examination, the USFMS scored lower (p=.05) than the USMS. When queried about this, USFMS indicated that this was the first time much of this information had been presented to them, while USMS reported that they had had prior exposure to some of the content of the clinical correlation lectures.

As part of the final evaluation, each student obtained a complete data base on an unknown patient and wrote a medical record using a problem oriented format, including a problem list, and assessment and plan for each problem. An independent physician rater reviewed all final write-ups submitted by students in the Classes of 1980 and 1981. USMS and USFMS did not differ on these ratings but the Class of 1981 scored significantly higher (p=.01) than the Class of 1980.

CONCLUSIONS

The USFMS as a group demonstrated deficiencies in the patient assessment skills of physical examination and medical interviewing at the onset of the six-week training program. In addition, there were significant differences between the skills of the USFMS and the USMS. At the completion of the training program, all differences between the USFMS and the USMS were non-significant except for knowledge of the content presented during the clinical correlation lectures. The only differences evidenced between the Classes of 1980 and 1981 were on the preceptor evaluation and the final patient write-up. Deficiencies in clinical skills were reduced by a brief but intensive training program characterized by specific mastery criteria.

REENTRY OF U.S. CITIZENS IN FOREIGN MEDICAL SCHOOLS INTO OUR MEDICAL EDUCATIONAL SYSTEM: PROBLEMS AND OUTCOMES

Barry Stimmel, M.D., Office of Academic Affairs,
Departments of Medicine and Medical Education,
Mount Sinai School of Medicine
of The City University of New York

United States citizens are pursuing the study of medicine in foreign medical schools (USFMS) in increasing numbers. There are several ways that these students can reenter the American medical educational system; however, until recently, little data have been made available with respect to the knowledge and skills of these students as compared to their counterparts in American medical schools. In addition, no data have been forthcoming comparing skills of USFMS accepted with advanced standing in American medical schools with those USFMS who remain in their foreign medical schools (FMS) for longer periods of time.

This discussion will focus on comparative data with respect to (1) students enrolled in American schools (USMS) and students accepted with advanced standing from foreign medical schools (COTRANS); (2) USMS and Fifth Pathway students (FPS); and (3) COTRANS students and FPS.

The themes that will be developed are: (1) Although a selected group of COTRANS students' National Board scores and subsequent clinical performance may not differ from USMS, certain initial deficiencies in skills can be identified that should be addressed in order to ensure a maximal performance on entering the clinical curriculum. (2) FPS who have two more years of training than COTRANS students still have similar deficiencies in certain skills. (3) Although Fifth Pathway programs (FPP) upgrade skills, the results, as judged by performance on state licensing examinations (FLEX), are quite variable, with pass rates still significantly below USMS. (4) Longitudinal studies of cohorts of USFMS reveal a considerable number either to have not yet taken the FLEX or to continue to fail it despite several attempts.

If the data presented in these selected cohorts are able to be projected for the entire pool of U.S. citizens studying medicine abroad, then an increasing number of students who have spent great time and energy in the study of medicine in FMS may not possess sufficient skills to qualify for the practice of medicine in this country. The "medical underground" of unlicensed, foreign-born physicians described several years ago may continue to exist, with the USFMS replacing the foreign national whose entry into this country has become more difficult since the passage of the Health Education Professional Assistance Act.

SYMPOSIUM

RESULTS OF A NATIONAL STUDY OF A COMPUTER SIMULATION (CASE) AND THE AMERICAN BOARD OF INTERNAL MEDICINE'S RECERTIFICATION EXAMINATION: THE MERIT PROJECT

Chairman: George D. Webster, M.D.

Participants: Lynn O. Langdon, M.A.
John A. Meskauskas, M.S.
John J. Norcini, B.A.
George D. Webster, M.D.

Discussants: Arthur S. Elstein, Ph.D.
Victor R. Neufeld, M.D.

Results of A National Study of A Computer Simulation (CASE) and The American Board of Internal Medicine's Recertification Examination: THE MERIT PROJECT

George D. Webster, M. D. (Organizer)
American Board of Internal Medicine
Philadelphia, Pennsylvania

Objective:

Since 1973 the American Board of Internal Medicine (ABIM), as part of its research efforts to develop relevant and effective measures of clinical competence, has supported the MERIT project (Model for Evaluation and Recertification by Individualized Testing) in conjunction with the Pacific Medical Center and Ohio State University.[1] The computer examination was developed from the CASE system. A national field test was carried out in January, 1978, designed to assess the measurement properties of the computer examination, to compare it with evaluation methods traditionally used for medical certification and recertification, and to assess the validity of each as compared to an external criterion of competence. The subjects in the study were 332 internists seeking Recertification six years or more following initial certification.

The symposium will present reports and discussion of data obtained in this study. Data will be presented on the scoring system developed for CASE, the interrelationships of these scores within the computer examination and with the written 1977 Recertification Examination, and the validity of each as inferred from weighted check-list evaluation of the subjects by peers.

---

[1] The Board wishes to acknowledge the major contributions to this project of William Harless, Ph. D., John Gamble, M. D., Nancy Farr and Marcia Zier of the Pacific Medical Center; Ruann Pengov , Ph. D., Mark Forman, Holly Fries and others of Ohio State University; Ralph O. Wallerstein, M. D., and the ABIM MERIT Committee.

An Overview of the Study

George D. Webster, M. D.

In early phases of this project, pilot studies of the CASE system (Computer-Aided Simulation of the Clinical Encounter--developed by Dr. William Harless at the University of Illinois and currently available at Ohio State University) indicated that internists' performance on CASE compared favorably with their performance as measured by chart audit. Therefore, the ABIM embarked on a study to compare this methodology with its written examination formats, and to assess the validity of each to an external criterion. Secondary objectives were to assess the effect of an internist's type of practice on examination performance, the effect of age, the subjects' perceptions of the computer examination, and technical questions relating to nation-wide administration.

Subjects: The subjects for this study were selected from the over 2,200 internists who had taken the ABIM's Recertification Examination in October, 1977. All had originally been certified in internal medicine at least six years prior to 1977. The candidates invited to participate were selected at random from two populations: those that indicated that greater than 50% of their practice consisted of general internal medicine; and those who were certified in cardiology as well as internal medicine. Approximately 60% (332) of those invited agreed to participate. For the purpose of addressing certain research questions, these participants were further broken down into subgroups according to their practice patterns: four of these groups (containing between 50 and 98 internists totalling 293) have been analyzed most extensively.

Measures: The data to be discussed are derived from three different sources:

.The Computer Examination which consisted of three CASES assigned according to the above groups. All examinees within a group interacted with the same CASES, but different groups were assigned different combinations of the 10 CASES that had been developed. Each of the CASES was developed, refined, and scored by a committee of expert Board members: details of the scoring system will be presented subsequently.

.The Recertification Examination was a one-day proctored written examination consisting of single best answer questions, multiple true-false questions, and linear Patient Management Problems.

.The external criterion selected for the study was the judgement of a participant's competence by his/her peers. The instrument used was a "weighted check-list" rating form, selected because studies in other fields indicated a high degree of interrater reliability and a high correlation with other external measures of performance. The check-lists consisted of behavioral statements written to reflect degrees of a variety of traits. Some were developed to reflect the computer examination scores and others the written examination scores. Weights were assigned to these statements based upon the judgements of 164 scaling judges who were internists who had previously been recertified by the ABIM. Rating forms were sent to the chiefs of medical service identified by each internist in the study, and the chief and two other internists nominated by the chief were asked to complete the form, identifying those statements in the form that were "very typical", "typical", "atypical", or "very atypical" of the internist being rated.

Two or three ratings were obtained on 60% of the 332 subjects. Factor analysis of the individual trait scales showed a strong common factor: i.e., the individual traits were highly intercorrelated. Furthermore, there were no significant differences in the mean ratings of the chiefs of service as compared to the peers. Therefore, all the scales and all available judgements were combined into a composite rating.

Several scoring systems for the rating scale were explored. Two, which differed mainly in the degree of confidence indicated by the raters' responses, were found to be equally reliable and only moderately correlated. Therefore, both have been used. The following data describe the external criterion measure:

	Mean Score of Candidates	Stand. Dev.	Reliability
Scoring System 1	2.96	1.01	.88
Scoring System 2	4.26	0.29	.92

Correlation of 1 with 2 = 0.52

Test Administration: The computer examination was administered at 15 test sites across the country linked to the OSU computer facility through the TYMNET communications network. Each site was staffed by a trained monitor and typists were available to assist internists who expressed the need. Participants were allowed 4 - 5 hours for the examination. No major technologic or logistical problems developed during the period of administration.

Internists' Perceptions: Each participant was asked to complete a questionnaire after finishing the computer examination. In general, they "agreed" that the CASES were representative of their practice, and were technically satisfactory. They judged the methodology was superior to MCQ's or PMP's in assessing their data gathering skills and equal to PMP's in assessing problem-solving skills, management and overall clinical competence.

This brief description of the history of the project, the subjects, the measures and the external criterion used will set the stage for the subsequent presentation of the results.

Lynn O. Langdon and John J. Norcini

AMERICAN BOARD OF INTERNAL MEDICINE

USE OF CASE AS AN EVALUATION MECHANISM FOR RECERTIFICATION

I. CASE DEVELOPMENT

A. Content: The content of each of the computer cases that was used in the January Recertification National Study was originally based on data from a real patient. The cases were revised and refined on the basis of extensive fieldtesting by internal medicine residents and practicing physicians. The complete content of each case was reviewed, revised and approved by an ABIM committee of internal medicine experts. A total of ten cases was utilized in the study; each participant's examination consisted of three CASEs.

B. Scoring: The criteria for several scoring parameters was established by the ABIM expert committee, fieldtested by practicing physicians, revised on the basis of fieldtest results and "key validation" alterations were made on the basis of ABIM Board members' performance.

The following eight scores were created for each case using criteria established for each case by the expert committee:

1. Data gathering of critical information in the history
2. Data gathering of critical information in the physical exam
3. Data gathering of critical information in the laboratory
4. Diagnostic correctness
5. Management correctness
6. Laboratory danger/discomfort
7. Laboratory efficiency (percentage of procedures/tests ordered that were considered acceptable or critical to perform)
8. CASE composite score (case average consisting of weighted subscores, 1-5)

II. CASE SCORE CHARACTERISTICS

A. Means and Standard Deviations:

Case	N	History	Physical	Lab	Diagnosis	Therapy	Case Average
Box	38	53(17)	89(14)	76(14)	88(22)	59(37)	73%
Cohen	109	67(20)	88(19)	79(15)	79(28)	71(43)	77%
Duncan	120	65(15)	86(11)	83(14)	95(11)	83(14)	82%
Gotrocks	126	61(19)	84(16)	57(17)	90(16)	41(29)	67%
King	106	75(15)	69(20)	76(20)	80(20)	71(29)	74%
Madison	86	56(17)	46(20)	29(27)	83(17)	57(21)	54%
Richardson	126	62(15)	71(12)	83(12)	91(16)	80(20)	77%
Thompson	85	75(15)	81(20)	81(18)	90(23)	83(21)	82%
Watson	108	76(20)	87(13)	95(13)	99(7)	80(21)	87%
White	88	60(16)	71(16)	54(10)	80(23)	78(28)	69%
Over-all average		65%	77%	71%	88%	70%	

These raw scores show considerable variation both within case and across case (range of total case difficulty: 54%-87%). Because of this variability the raw scores were standardized for analyses. It is important to note that the sample that took each case was not constant so the standardization of each case is based on different populations. For this reason, only cases that were taken by large numbers of examinees were included in subsequent analysis.

The intercorrelation matrix of simulation scores utilizing the standardized average of each of 7 scores for the 3 cases in each examination (n = 289) is shown below:

	Hx	PE	Lab	Eff	D/D	Dx	Rx
Hx critical concepts	1.0	.38	.42	-.04	-.10	.38	.20
PE critical concepts		1.0	.44	-.06	.04	.24	.24
Lab critical concepts			1.0	-.01	-.04	.37	.34
Lab efficiency				1.0	-.44	.08	-.01
Lab danger/discomfort					1.0	-.09	-.11
Dx correctness						1.0	.47
Rx correctness							1.0

There is a wide diversity in patterns and extent of intercorrelation suggesting that case content may be a significant factor in performance.

B. Reliability: The computer case reliabilities* based on standardized composite scores for each case (30-60 minutes of testing) ranged from .37 - .74 with a mean of .62. Comparison of these figures to reliabilities of ABIM written linear patient management problems shows similar results; the range of six PMP reliabilities* from the 1977 Recertifying Examination was .51 - .72 with a mean of .642. However the testing time required for a written PMP is shorter than for a computer case.

The reliabilities of computer composite scores** - combining the three cases in each participants' examination - were:

Group	N	Reliability	Cases
I	95	.78	Cohen, King, Watson
II	49	.79	Duncan, Gotrocks, Richardson
III	71	.70	Duncan, Thompson, White
IV	74	.74	Gotrocks, Madison, Richardson

*Coefficient-alpha
**Mosier reliability

C. Effects of Individual Differences

The effect of age (inferred by the year in which the medical degree was granted) was considered and as can be seen in the table below, individuals who completed their medical degree more recently tend to do better on all test formats studied, but the effect is less for the two simulation models than for multiple choice questions.

Correlations Between Year of MD Degree
And CASE, Recertification Exams

	Year of MD Degree
Total Written Recertification	.50
One best answer	.49
Multiple true/false	.44
PMP	.37
CASE Component Score	.35

All of the correlation coefficients are statistically significant at $P < .05$.

The question of effect of specialization was considered in relationship to performance on the computer simulations only - and only three cases were appropriate to the analysis. Duncan and Richardson were cardiology cases, Gotrocks was a gastroenterology case. It was predicted that individuals who consider themselves subspecialists in an area would perform better on cases in that area.

In all three comparisons that were made, the predicted trend was apparent but the only improvement that was statistically significant ($P < .01$) was for the gastroenterology area of specialization; however a less dramatic improvement in cardiology could be explained by the fact that cardiology is a common area in general internal medicine and may be generally more familiar to internists.

VALIDITY, PMP/CASE REDUNDANCY, AND SOME COMMENTS ABOUT THE STRUCTURE OF RELATIONSHIPS AMONG MEASURES

John A. Meskauskas, M.S. and John J. Norcini, B.A.
American Board Of Internal Medicine

This portion of the symposium will address three major topics: the structure of relationships among measures and their possible implication for a theory of clinical competence; PMP/CASE redundancy or overlap; the validity of MCQ's PMP's, and CASE as assessed by a weighted checklist filled out by the study participants' colleagues.

The Relationships Among Measures And Their Theoretical Implications

To assess whether CASE scores reflected competence, 17 Board members were asked to rate copies of CASE interactions (5 raters per interaction) in terms of the overall competence exhibited as well as History-taking, physical findings, lab studies, avoidance of danger/risk, diagnosis and management. Ratings on overall competence were highly reliable: The range was .71 to .98 across 10 simulations. As the intercorrelations among the seven rated variables were generally high (80's) within each case, it was clear that the overall rating tapped all the variance to be had. Furthermore, the overall mean Board rating correlated .80 with the CASE computer-generated overall score. Thus CASE scores closely reflect Board judgments of overall competence, but the mechanism by which this occurs is puzzling because the component variables of the two overall scores have a very different structure. All of the ratings are highly intercorrelated, supporting the notion that, at least when performance on a single case is assessed, competence may be a unitary trait. The previous paper, however, showed us that the correlations among CASE scores are not high, which would imply that we should view competence as multivariate. Since unreliability could account for the low correlations, a multitrait-multimethod analysis (Campbell and Fiske), which looks at relative degrees of relationship and hence in essence controls for unreliability, was performed. The five component scores of each of the three CASE patients given to a sub-group (N=95) were used in this analysis. The across-case correlations (Table, 1 first column) for a single trait (e.g., history in CASE 1 versus history

Table 1

Trait	Average correlation for the given trait across different case	Average correlation for the given trait with other traits in the same case
History-taking	.30	.21
Physical exam	.30	.19
Laboratory	.03	.25
Diagnosis	.25	.23
Therapy	.02	.21

in CASE 2) are low and no greater than the within - case correlations across traits (e.g., history with physical, lab, etc.). Thus while CASE clearly measures overall competence it is not useful to think that it measures that competence by assessing traits such as history-taking, diagnosis, etc., which are often viewed as generalizable traits across different clinical problems. We find no support for such a notion, so some other conceptual framework is needed.

Are Patient Management Problems (PMP's) And CASE Redundant?

Since PMP's and CASE are two different forms of simulation problems, one can ask whether they measure the same traits. To test this, two CASE's were developed from Recertification PMP patients and included in the computer examination. We thus had two "content-matched" pairs of measurements although the extent of matching was somewhat limited by differences in the formats. For example, the computer simulation presented a much richer history than the PMP's but it was self-generated by the examinees whereas in the PMP it was given. Total scores between the content-matched pairs of cases were correlated; Table 2 presents these. It was hypothesized that, if PMP's and CASE measured the same aspects (were redundant) then the correlations between content-matched PMP and CASE scores (same content, different method: condition A) would be high, or at least higher than the correlations among PMP or CASE scores (same method, different content: conditions B and C). The correlations in conditions A, B, and C should each in turn be higher than the PMP versus CASE score correlations (different content, different method: condition D).

Table 2

Condition	Average Correlation
A. Correlations between content-matched PMP and CASE score	.16
B. Correlations among PMP scores	.28
C. Correlations among CASE scores	.35
D. Cross-method correlations - unmatched PMP versus CASE scores	.18

Since the correlations in condition A are no higher than those of D, it is clear that the method by which that similar content is assessed (PMP or CASE) makes a difference. This conclusion of non-redundancy is reinforced by the higher correlations of conditions B and C. These suggest a small but larger amount of within-method consistency of performance.

Validity Of Instruments As Assessed By A Weighted Checklist Criterion

To assess validity, the weighted checklist described previously was developed containing 150 items. Scales were combined via factor analysis, resulting in highly reliable total scores (alpha reliability was .92 for the scoring method reported here). The validity of each variable against this criterion was assessed (Table 3) both singly, (r) and as part of a stepwise multiple regression (R, R^2, etc.). The univariate correlations ranged between .28 and .30, suggesting that each question format is valid and about equally so. The stepwise regression shows that CASE adds more variance than any of the other variables entered after the first step, but that the absolute amount added to the currently-used variables (.026, or 2.6% of the variance that could possibly be predicted) is not large.

Table 3

Step	Variable Added	r	R	R^2	increase in R^2
1	One-best answer Question	.28	.285	.081	-
2	True-false Questions	.30	.319	.102	.021(26%)
3	PMP	.28	.341	.116	.014(14%)
4	CASE	.30	.377	.142	.026(22%)

In relative terms, since the currently used question formats account for 11.6% of the variance in the criterion, and the addition of CASE raises that to 14.2% the increase in validity (22%) is sizeable.

References

Campbell, D.T., and Fiske, D.W. Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. *Psychological Bulletin*, 1959, *56*, 81-105.

Elstein, A.S., Shulman, L.S. and Sprafka, S.A. *Medical Problem Solving: An Analysis of Clinical Reasoning*. Cambridge, Mass: Harvard University Press, 1978.

La Duca, A. The structure of Competence in Health Professions. Paper presented at Annual meeting of American Educational Research Association., San Francisco, April 1979.

SYMPOSIUM

PERSPECTIVES ON HUMANISTIC MEDICAL EDUCATION

Chairman:	Richard Fletcher, Ph.D.
Participants:	Robert T. Manning, M.D. Cindy Taylor Gerald Holman, M.D. Ronald Jordan, Ph.D.

Purposes Of The Symposium

1. To review major developments in programming in Humanistic Medical Education with a view toward:
   a. Highlighting some of the research approaches to evaluation of such programs and their findings.
   b. Pointing to the most promising lines of development for such programs.
   c. Preparing for more interchange among programs so that the more promising efforts can serve as models.

Robert T. Manning, M. D., Professor of Medicine, Director of Internal Medicine Education, Wesley Medical Center, University of Kansas Medical School,Wichita

HUMANISTIC MEDICINE-A PERSONAL PERSPECTIVE

Humanism, or the humanistic approach as applied to clinical practice, refers most often to interaction with a sick individual as a person, not as a disease encased in a body. It is best summed up in Francis Peabody's statement "Patient care begins with caring for the patient" to which I would add parenthetically, caring for the patient as a unique whole human being.

Caring for patients as persons requires constant vigilance on the part of the physician less he is diverted by a fascination with disease (an abstraction) or technology away from consideration of illness (the patient's reality). Faced always with a degree of uncertainty in the clinical and diagnostic processes, the physician readily and understandably seeks "the scientific" information, the facts with which his entire preparatory career has disciplined him to be most comfortable with. Compounding the problems for the physician is the inhuman educational process that leads to his professional role. Throughout his training he may have seldom been treated as a person; seldom been valued; seldom felt he was in a caring environment. This loss of a sense of personal worth is certainly and decisively a contribution to many of the dehumanizing aspects of medical care, particularly for hospitalized patients.

Humanism works in the classroom, as demonstrated by David Aspy's study, The Element of a Caring Relationship. Empathy, Congruence and Positive Regard, as formulated by Carl Rogers, are as applicable in the classroom or at the bedside as in home. If a medical student, or a resident, or nurse, or physician has no sense of personal worth, or whose empathy has been suppressed and is unable to be congruent in behavior because of external demanding rigidity in training, it should again be little wonder that expression of these qualities toward others is impaired if not fatally afflicted.

I believe, therefore, that humanism at the bedside is fostered best by a humanistic approach to education and training and by a system that values the student as a person and positively rewards the care of the sick at least as much as the diagnosis of disease.

My personal experience with the human dimensions in medical education program leads me to the inescapable conclusion that its goals are realistic and obtainable. They certainly work for me, and my observations of the students at Eastern Virginia Medical School supports my beliefs. The best evidence is to ask any of the students "What is your life like? Do you feel cared for and valued as a person?" My non-randomized uncontrolled but perspective study tells me they have more positive human sensitivity to themselves and others than medical students elsewhere. Will they be better physicians because of these efforts? I suggest we must wait until their careers have matured and ask again and perhaps more importantly, as the wives, children and patients for whom they care.

Representative of the Humanistic Medicine Task Force of the American Medical Student Association, Cindy Taylor, MS 81, Medical College of Pennsylvania

THE HUMANISTIC MEDICINE TASK FORCE

The Humanistic Medicine Task Force of the American Medical Student Association was conceived at a workshop in 1976 jointly sponsored by AMSA and the San Francisco-based Institute for the Study of Humanistic Medicine. The Institute was a group of health professionals created in response to a perceived loss of the importance of the human side of both physicians and patients in this country. Workshop participants explored philosophies and techniques from the "human potential movement" and healing modalities from various cultures. They established a set of principles to address the human factors in healing.

These principles aim at the recognition that a patient cannot be seen simply as a disease, nor a health professional solely as an expert in medical technology; the full healing potential of their relationship depends upon the quality of their interaction as whole human beings. Health is seen as a dynamic equilibrium among the factors that comprise and affect a person (body, mind, emotions, spirit, environment, and other individuals), and when that equilibrium is disturbed, regaining it must be a collaboration designed to evoke the individual's own human and biological resources. Part of this process must be an understanding that illness has a unique meaning in the context of each individual's life, and a sensitive health professional can help a person use the experience creatively to a positive result which maintains the integrity and reaffirms the dignity of the person.

The Task Force developed as a group of students who made a committment to continue to work through AMSA to infuse these principles into the philosophy and practice of medicine. In addition, these students became concerned about the dehumanization inherent in medical training, and addressed themselves to change at all levels of the process.

Most of the Task Force's activities have been experiential workshops designed to facilitate insight into the experiences of health and illness, the roles played by patients and practitioners, and the problems of maintaining one's humanity during the medical training process. Workshops have been held at two national conventions, fall and spring regional workshops, and local programs around the country. Examples of some of the themes explored have been "surviving medical school", "communities and care", "alternate forms of health care", "experiencing the self in health and illness", and "the creative experience of illness". In addition, several cities and regions have produced local bulletins and newsletters, and a resource compendium for workshops has been compiled.

ORGANIZATION

The Humanistic Medicine Task Force has a national coordinator whose function is to maintain an open communications network between the national organization and regional and interest area coordinators, and correspondence and liaison on behalf of the TF with outside groups with related interests. In addition, the national coordinator helps to channel requests for assistance with local programming into the proper resources and has the final responsibility for Convention and other national projects.

A newsletter coordinator reports on projects and serves as financial advisor to the national coordinator. In addition, there is a regional coordinator for each area who is responsible for coordinating the efforts among individual chapters in that area, and programming for regional workshops.

This year the Task Force has identified five areas of interest for discrete program planning. Each interest area has a coordinator whose function is to gather information pertaining to the Task Force's goals in this area, and help to funnel program activities relating to it. These coordinators will communicate with regional coordinators regarding regional projects in their area, and serve as special consultants to regions for local workshops. The five interest areas are:

1. Curriculum development
2. Medical student support and well-being
3. Political and social issues in health care
4. Self care and alternative forms of healing
5. The experiences of health and illness, and the nature of related intrapsychic and interpersonal interactions

Regional coordinators will decide with their local chapters which of these interest areas they wish to explore, and develop their resources accordingly.

LONG-TERM GOALS

It is the hope of the HMTF that, by a combination of four approaches, significant progress may be made in orienting medical training and practice toward the whole person.

1. Workshops and Training Experiences for Students
    a. Experiential workshops in each of the interest areas
    b. Training workshops to develop desirable skills in the largest possible number of students
    c. Externships and practical experiences in Humanistic Medicine
2. Education of Medical Educators to nurture the development of approaches which help integrate emotions, social, and spiritual dimensions into the training of physicians by changing the consciousness of those who serve as role-models.
3. Modification of Medical School Curricula to incorporate humanistic dimensions of learning
    a. Incorporation of interpersonal and behavioral science skills into medical training from the beginning, including sensitive interviewing and examining techniques and the physician's role as health advocate.

377

   b. Incorporation of the Humanities into the standard preclinical curriculum, including such courses as bioethics, lifecycle, medicine and literature, medicine and philosophy, history of medicine, and images of healers.

   c. Integration into curricula of opportunities for exploration of feelings and reactions to situations encountered in medical training, such as dealing with experiences of loss and death.

   d. Modification of teaching methods to include less traditional approaches which encourage real empathy with the patient, such as simulation techniques, videotape and peer feedback on rapport skills, and the use of surrogate patients for interview and exam skill feedback.

4. Student and Physician Well-being and Support; preventive approaches to the problems of impaired physicians.

   a. Development of peer and professional counseling networks for students and housestaff, such as support groups, "buddy" systems, weekend retreats and orientation programs. This would also include development of special support for the unique stresses on medical people (feelings about death; accommodating lifestyles to the demands of a medical schedule, etc.) and for groups of medical people with special needs (such as women, gay people, minorities, handicapped people, etc.).

We hope that the HMTF will function as a clearinghouse for ideas and projects addressing these concerns and assisting students interested in working for change in their own areas.

Gerald Holman, M. D., Dean, Eastern Virginia Medical School, Norfolk, Virginia

CAN MEDICAL STUDENTS LEARN TO BECOME HUMANISTIC COMPASSIONATE PHYSICIANS?

Introduction

In the past two decades there has been an increasing concern by society about the technological emphasis in medical education. Indeed, some believe that medical schools are actually producing physicians who are anti-human; that is to say that the physician graduate is more concerned with things and events than people and focuses on those elements rather than the person as an individual. Baldwin (1) has stated that medical schools were producing anti-human, anti-intellectual and anti-scientific physicians. Further, Rezler (2) in reviewing the literature reported that medical school increases cynicism in medical students.

Historical Background

The Eastern Virginia Medical School (EVMS), a three year medical school, located in Norfolk, Virginia accepted its first students in the fall of 1973. The founding dean stated that one of the emphases of the medical school was to produce scientifically able compassionate physicians.

In February, 1975, upon requests of several faculty of the Department of Psychiatry and Behavioral Science, the second dean agreed to send several faculty and students to the Human Dimensions in Medical Education (HDME) workshops sponsored by the Center for the Study of the Person in La Jolla, California. It was thought that such experiences would enhance the commitments of the school to the education of compassionate physicians. This program was chosen to determine what contribution it might make in achieving the stated objective of graduating compassionate physicians.

Faculty and students returned highly enthusiastic. Based on this preliminary response and with grant support from the Kellogg Foundation, a multifaceted human dimensions program was undertaken in the fall of 1975 as a model program.

The program included opportunities for faculty, students and administrators to attend four and occasionally ten day HDME workshops. An HDME steering committee of faculty and students was established to oversee and evaluate the program, and an ongoing consultant of the HDME program in California guided this steering committee. The evaluation system was designed by the Center for Educational Development. The incoming class of 1976 became the first class to participate in the full program.

Essential elements include:

Parent and student participation in student orientation day and the introduction to an approach to patient interviewing on this first day of medical school.

A faculty and student retreat including spouses in the mountains of Virginia for three and a half days. This retreat takes place immediately following orientation day and consists of small group discussions about students and faculty concerns. This is held early in the medical school experience as the first official medical school activity, to achieve high student visibility.

Continuing small group activities both as a formal part of the curriculum and informally as support group for students.

An additional two day retreat for students, faculty and spouses again away from Norfolk approximately one-half way through the first 15-month phase of the curriculum.

The major purpose of this program was an attempt to develop an open supportive communication system between students themselves as well as between students and faculty. An additional objective was to produce a sense of "community" within the medical school family.

Over the past three years the program has changed in some measure. HDME has become but one component in a broader Human Values in Medical Education program. This program consists of the following elements:

The HDME activities
Bio-medical ethics
An expanded student advisory program
A humanities interdisciplinary educational program
Emminent speakers program
Interaction with the interpersonal skill component of the curriculum

Six different internal evaluations have been compiled, including:

1. Perception of the learning environment (semantic differential)
2. Review of annual retreats
3. Small group effectiveness
4. Semi-annual retreat reveiw
5. Interpersonal skills curriculum assessment
6. Follow-up of EVMS graduates

Semantic Differential

As described more completely in another presentation of this symposium, a semantic differential was constructed to measure the educational climate. Medical students in Germany provided a source for comparing differing educational settings. Both American and German students were asked to rate the educational environment "As It Should Be" and As It Is".

Annual Retreat

The annual school retreat held in June, 1978, was reviewed with respect to its effectiveness in improving interpersonal relationship skills, and this documented several changes. There were significant increases in interpersonal skills and a decrease in anxiety when pre- and post-tests were compared. Attitudes towards the medical school and the interpersonal relationships of participants were also substantively improved. Details of this evaluation will be provided during the symposium but examples of individual interpersonal skills changes include an improvement in the ability to listen to others in an understanding way; a change in the awareness of the feelings of others; an increased tendency to trust others; a reaction to the opposing opinions of others; and an increase in peace of mind.

The effect of this particular retreat was not as great as some had anticipated and has led to redefinition in part of the structure of future retreats.

Follow-up evaluation by telephone questionnaire of students who have graduated from EVMS and who are now at least one year into their graduate medical education experiences suggest that the humanistic program has been effective at least in the students perception as they believe they are more caring of patients and more concerned about each person's welfare rather than focusing on diseases and events. Whether this emphasis will be maintained after several years of clinical practice awaits further evaluation.

Summary

The substantial emphasis given to human interaction at EVMS as the basis for the future practice of medicine has developed an open communication system and several support systems to assist the student in coping with the difficulties of medical education.

It is our belief that medical students can learn to become caring compassionate physicians and that an open communication system between faculty and students will facilitate the achievement of this objective.

Bibliography

1. Baldwin, D. C. Can Holistic Medicine Be Taught In Medical School. American Holistic Medical Association, February, 1979.

2. Rezler, A. G. Attitude Changes During Medical School: A Review Of the Literature. Journal of Medical Education, 49:1023-1030, 1974.

381

Ronald D. Jordan, Ph.D., Director, Center for Educational Development,
Eastern Virginia Medical School, Norfolk, Virginia

SEMANTIC DIFFERENTIAL AS A MEASURE OF ATTITUDES TOWARD THE MEDICAL SCHOOL ENVIRONMENT

Studies on attitude change in medical school tend to indicate that students develop increased cynicism during their medical training. Eron, for example, reported in 1955 that humanitarianism in medical students decreases during their medical education. Eron repeated the observation as the result of a follow-up study several years later, and several additional studies, such as those conducted by Kepler and Saslow, and Gordon and Mensh are consistent with the findings of Eron. In her recent summary of the literature, including studies with conflicting conclusions, Rezler reports that medical school does, in fact, cause an increase in cynicism; furthermore, special programs do little to counteract this trend unless supported and reinforced by the total medical school environment.

Assessment of the medical school environment, therefore, is a necessary component of any systematic effort to stem or reverse the development of negative attitudes. This is particularly true at institutions where there is a strong emphasis on the development of humanistic, compassionate physicians.

In order to validate an instrument that would be sensitive to attitude changes over time during medical school, it was felt that it should be shown to have sensitivity to differences between two educational environments at the same time. One American and one German medical school were chosen as representatives of two contrasting educational environments. The American school is small, community based and supported, and unaffiliated with a university. A major goal is the development of physicians who will specialize in primary care and whose practice of medicine will be humanistic and compassionate. It was the first of several medical schools across the country to be funded for programs in Human Dimensions in Medical Education. The German school is very large, federally funded, university-based and has no stated goals or programs pertaining to the development of humanistic physicians.

A semantic differential was constructed containing thirty-five bipolar descriptors deemed important to assessing the total educational environment. Descriptors were placed on either side of a seven point scale in a random fashion, so that descriptors judged positive by the experimentor occurred on the lower end of the scale at the same rate as descriptors judged negative. The thirty-five bipolar descriptors were placed on two different forms. One form was designed to assess the concept of the Educational Environment As It Should Be, and the other to assess the Educational Environment As It Is. The forms were distributed to 48 American medical students and 102 German medical students, all of whom were at least mid way through their pre-clinical programs. All students received either a Should Be or an As It Is form, and were asked to identify the point on the seven point scale which reflected their opinion of the educational environment. Thus, a rating of 7 or of 1 would be the strongest, and a rating of 4 would correspond to a neutral or no opinion. Prior to analysis of the data, those scales on which the positive descriptors were on the lower end were reversed so that higher scores would reflect more positive assessments.

Results

A 2 x 2 analysis of variance of overall mean scores on the semantic differential showed that American students scored significantly higher than German students ($p < .001$), and that there were significantly higher responses on the Should Be form than on the As It Is form ($p < .001$). A significant interaction ($p < .001$) was obtained, due primarily to the fact that German students were substantially more negative than Americans about the environment As It Is. In contrast, the two groups were almost identical on what the environment Should Be. Scheffe' *post hoc* analysis supports these conclusions by showing a statistically significant difference on the As It Is form, but not on the Should Be form.

Discussion and Conclusion

The results show the semantic differential to be sensitive to differing educational environments. The differences between the German and American students were not only statistically significant, but also quite meaningful. It was not uncommon on the As It Is form to find scales on which the American students felt very positive (a mean score above 5.0), and in which the German students felt negative (mean score of only 2.5 - 3.4). Thus, it was possible to characterize the American students' environment as relaxed, good, warm, flexible, healthy, caring and humanizing. The German students' environment, on the other hand, was shown to be tense, bad, cool, rigid, sick, indifferent and dehumanizing.

The results show the semantic differential as a valid measure which does not require so much time to fill out as to be a "dehumanizing" experience in itself for busy medical students. The ease of administration commends the instrument to repeated use during the curriculum, e.g., before admission, after the first major exam, or after each clinical rotation. Thus, it can be used to determine those programs which contribute to or help to stem the trend toward cynical attitudes shown by research to occur in many medical education programs.

SYMPOSIUM

RESEARCH ON WOMEN PHYSICIANS: RECENT EMPIRICAL FINDINGS AND AREAS FOR FUTURE INVESTIGATION

Chairman:	Dorothy Rosenthal Mandelbaum, Ph.D.
Participants:	Lillian Kaufman Cartwright, Ph.D. Sandra L. Chaff Marilyn Heins, M.D. Dorothy Rosenthal Mandelbaum, Ph.D.

384

SYMPOSIUM

RESEARCH ON WOMEN PHYSICIANS: RECENT EMPIRICAL FINDINGS AND AREAS FOR FUTURE RESEARCH

Organizer: Dorothy Rosenthal Mandelbaum, Ph.D.
Rutgers, The State University
Camden, New Jersey 08102

Participants: Lillian Kaufman Cartwright, Ph.D.
Dean for Academic Affairs
California School of Professional Psychology,
Berkeley, California 94704

Sandra L. Chaff, M.A., M.L.S.
Director, Archives and Special Collections of Women in Medicine
The Medical College of Pennsylvania
Philadelphia, Pennsylvania 19129

Marilyn Heins, M.D.
Wayne State University
Detroit, Michigan 48201

Dorothy Rosenthal Mandelbaum, Ph.D.

Objectives: The two major objectives of the symposium are:

(1) to present recent empirical findings from three ongoing studies of women physicians designed to explain career behaviors, and

(2) to introduce the Archives and Special Collections on Women in Medicine and areas for further research.

Research on Women Physicians: Recent Empirical Findings and Areas for Future Research

Dorothy Rosenthal Mandelbaum, Ph.D.
Rutgers, The State University
Camden, N.J. 08102

The literature on women physicians is rich with biographical and anecdotal material which touch inevitably upon the relation between career behavior and adult female roles. In recent years these primarily descriptive records have been complemented by a growing number of empirical studies designed to analyze the reasons behind the career behaviors of women physicians. One purpose of this symposium is to share findings from three ongoing studies across the country which have been designed to analyze and to explain the behaviors of women physicians. A second purpose is to introduce an archive on women physicians along with suggestions for future lines of investigation.

Lillian Kaufman Cartwright has been following a sample of former medical students at the University of California Medical School in San Francisco since 1964. She will describe three basic styles of integrating work and intimacy adopted by the women of her sample. Marilyn Heins, who has been studying the productivity of women physicians in the Detroit area, will discuss factors which affect medical work records. Dorothy Rosenthal Mandelbaum will share recent findings of her longitudinal study of the career persistence of women physicians in the Philadelphia area. Sandra L. Chaff will describe the Archives and Special Collections on Women in Medicine at the Medical College of Pennsylvania and incate areas for further study.

The Integrative Challenges of the Young Woman Doctor: A Step Beyond the Barriers

Lillian Kaufman Cartwright, Ph.D.
Dean for Academic Affairs
California School of Professional Psychology
Berkeley, Cal. 94704

The current research is part of a more comprehensive 27-year study of approximately 1,100 male and female physicians, directed by Harrison G. Gough at the Institute of Personality Assessment and Research, Berkeley. To better understand the career development, satisfactions, and problems of women doctors, I followed a sample of young women longitudinally using a variety of methods including personal interviews, psychological inventories, and archival records. These 58 women were first interviewed when they were students at the University of California, San Francisco, School of Medicine. They comprised 98% of the geographically available females from the entering classes 1964-67.

On entry to medical school the women as a group were advantaged compared to other young women--they came from small families, their fathers were more educated, their primary family was more stable, and they were very independent and well-functioning as measured by personality inventories. Although they entered medical school at a time when there was little social support for their ambitions, their personal milieu, i.e., family and friends, were encouraging and reinforcing.

The follow-up interviews were conducted in 1975-77, with 51 of the original 58 women. The women now in their early 30's are either in the last stages of their residency training or in the beginning stages of practicing medicine. Among other tasks, the doctors were asked to rate their satisfaction with medicine as a career and also rate the amount of ease or difficulty they had in integrating their professional lives and their personal lives (i.e., their role as a woman, wife, and mother). Other studies on women physicians assert that women in medicine, particularly in the early stages of their careers, are under considerable role strain because of value conflicts, a nonsupportive professional climate, as well as the more obvious burdens of trying to integrate a taxing profession with a personal life.

Career satisfaction was extremely high in the sample with about 88% reporting they are satisfied or very satisfied with their profession. However, over half experienced at least a moderate amount of strain in integrating professional and sex roles. Strain came from having too much to do--the problem of juggling multiple responsibilities was apparent in the significant correlation between age, number of children, and strain. Personality characteristics, including a high degree of tolerance and lack of rigidity, as well as general intellectual resourcefulness and independence mitigated strain and was associated with "Role Harmony."

In order to identify the factors leading to satisfying combinations of work and intimacy, the 19 women who rated themselves high on both Career Satisfaction and Role Harmony were compared with the rest of the physicians. Individual differences were apparent: the women high on both ratings were exceptionally confident, intellectually resourceful, and tolerant. These 19 doctors tended to have clear priorities, with either family or work coming first. Overall, their current level of personality adjustment was superior to the others, who as a group were extremely well-functioning.

In the symposium I will depict three women who illustrate three basic styles of integrating work and intimacy. These styles are characterized as (1) "Medicine is my lust," (2) "My family comes first," and (3) "Super-woman." The costs and benefits of each styles will be evaluated. Stereotypic views and myths concerning each style will be dispelled.

Women Physicians: An Analysis of Factors Involved in Productivity

Marilyn Heins, M.D.
Joanne Hendricks
Wayne State University

Research about women physicians with and without training and practice interruptions has uncovered differences in personality characteristics and acceptance of conventional gender role definitions (1, 2). What differences are there between women physicians who have a "perfect" productivity record and those who do not? How do women physicians with and without children differ from each other when they both have perfect medical work records?

The Detroit area physician study which collected interview data from random samples of women and men physicians found that 59% of the women had a perfect medical work ratio (MWR equals full-time equated months in medical work since medical school graduation divided by total months since medical school graduation) (3). This paper reports differences between groups of women physicians in our sample (those who reported interruptions or worked part time). In addition, we discuss differences within the group with perfect medical productivity controlling for the presence or absence of children.

Women with perfect work records were more likely to be board certified, less likely to be married, had fewer children and performed fewer chores at home than did other women physicians in the sample. They reported more Career-Priority and fewer Family-Priority decisions as well as less conflict about these decisions. They were less likely to mention instances when other people and life circumstances altered their career plans ($p < .001$), were less likely to make specialty choices to accommodate gender role expectations ($p < .01$), and more likely to believe that careers are as important as husband and family to a woman. Those with perfect productivity were more motivated to become physicians as a result of exposure to illness and death ($p < .01$) and less likely to mention the prestige of medicine as a reason for career choice ($p < .01$) than were women with less than perfect work ratios. Life satisfaction ratings in the two groups of women physicians were not greatly different, but assessments of the use of skills and talents were higher ($p < .05$) for full productivity women.

Among the fully productive women, the women without children were younger and largely never (or not yet) married. These nonmothers were more likely to be a first or only child and less likely to mention encouragement of others either as a reason for career choice or in connection with career-family decision-making than were the women with children. The mothers did more chores at home, had more domestic help, reported more instances when their career affected family decisions, and experienced more career-family conflicts than did the productive women without children. The older women with children also tended ($\underline{p} < .07$) to rate their use of skills higher than those without children.

These findings about highly productive women physicians make it clear that the presence of children does not by itself lead to work interruptions. The data also lend support to the importance of gender-role beliefs in making career-family choices.

---

(1) Cartwright, L. K. Continuity and Non-Continuity in the Careers of a Sample of Young Women Physicians. JAMWA 32:316-321, 1977.

(2) Mandelbaum, D. R. Toward an Understanding of the Career Persistence of Women Physicians. JAMWA 31:314-324, 1976.

(3) Heins, M., Smock, S., Jacobs, J., et al. Productivity of Women and Men Physicians. JAMA 236:1961-1964, 1976.

The Career Persistence of Women Physicians: Five Years Later

Dorothy Rosenthal Mandelbaum, Ph.D.
Rutgers, The State University
Camden, N.J. 08102

The purpose of this paper is to explain changes in the medical, marital, and maternal role behaviors and attitudes of a sample of licensed women physicians who were studied initially in 1973-74 and again in 1978-79. The original study suggested that career persistence in licensed women physicians is an expression of differences in psychological developments between two groups of career-involved women. The longitudinal study was designed to further test the basic assumptions of the initial investigation.

Original Study

The original sample of 71 licensed women physicians was dichotomized into Persister ($\underline{N} = 40$) and Nonpersister ($\underline{N} = 31$) subsamples. The data were collected by means of a mailed questionnaire, taped interview, and a personality inventory.

The results indicated that the career persistence of women physicians has three possible explanations: (1) those who persist do so because they experienced a more difficult path to their career goal; (2) such persistence is often due to an early career interest development which imbeds the goal; and (3) those who tend to withdraw are more traditionally feminine in psychological development. Persisters (P's) tended to originate from families of lower social class origin which offered less financial and psychological support to their daughters' medical career goals. However their career goals tended to be strengthened by their characteristically early-expressed wish for a medical career coupled with superego-oriented motivation. After licensing the P's were less likely to marry or to mother. When they did assume either role, they tended to maintain the primacy of their careers.

NP's developed in families of higher social class in which they found more acceptance and support for their medical career goals. These women tended to develop more traditionally feminine personalities because of closer father-daughter relationships. Therefore, these women were more likely to integrate both traditional and professional roles equivalently and tended to withdraw from their careers on occasion for family reasons.

Longitudinal Study

Five years later 57 of the 65 living subjects (30 P's and 27 NP's) responded to a mailed questionnaire and personality inventory. The results of this study tended to confirm the initial findings.

Physician Role. Most striking was the increased investment in the physician's role by both groups. As expected no P withdrew. Only one NP withdrew and seven returned to work. Both groups tended to work longer hours weekly than they had in 1973-74. Six received board certification (four P's and 2 NP's). P's evaluated the physician's role even higher than they had earlier and fewer considered leaving medicine. The NP's, on the other hand, were less satisfied with their own self-expectations as physicians than they had been in 1973-74.

Marital Role. As expected the P's still gave precedence to their careers. Marital-career strains tended to be resolved in favor of the career. A higher proportion of separation/divorce occurred among P's (4/14) than among NP's (2/27). NP's reported that their husbands were more pleased with their medical careers, thereby dissolving an earlier difference between groups.

Maternal Role. NP's still invested more heavily in their maternal roles than P's. They averaged more children and were more actively involved in their rearing. Yet they became less satisfied with their maternal role: they were more concerned with their children's rearing and were less satisfied as mothers. They further perceived that their children grew less positive to their physician role than P children, a reversal of the 1973-74 findings.

Women. The earlier difference between P and NP on their perceptions of themselves as women dissolved.

Changes in the attitudes and perceptions of the NP's may be explained by the return to work of seven group members. On the positive side these women reported increased ease with other women and greater acceptance by their husbands. On the negative side they were more unhappy with their maternal roles and were somewhat more impatient with their professional achievements. Nevertheless only four NP's (and no P) found the five-year interim a period of dissatisfaction. On the whole, both groups reported the period one of personal growth (N = 11) or acceptance of static life situations (N = 9).

We may conclude that the basic assumptions of the original study held true five years later: P's continued to give their medical careers first priority in their lives and NP's were still more traditionally feminine in attitude and behavior. What is quite striking, however, is that overt differences in career behaviors narrowed with time.

Research on Women Physicians
The Data Base (with Special Reference to the Medical College of Pennsylvania's Women in Medicine Collections), and Implications for Future Projects

Sandra L. Chaff
Director, Archives and Special Collections on Women in Medicine
The Medical College of Pennsylvania
Philadelphia, Pa. 19129

One of the most comprehensive collections of material on women physicians is the Archives and Special Collections on Women in Medicine at the Medical College of Pennsylvania in Philadelphia. The Medical College of Pennsylvania, founded in 1850 as the Female Medical College of Pennsylvania, was the first and is the only extant medical school regularly organized for the education of women to be physicians. The archives (or noncurrent records of the College and and Hospital) consists of Board of Corporators and faculty minutes dating back to 1850, financial records, student and alumnae materials, and clinical case books (particularly from the medical and obstetrical clinics).

In addition to the records of the Institution, the Archives and Special Collections houses the records of other medical women's institutions and organizations, such as women's hospitals, women's medical schools, and women's medical societies. Other special collections include the papers of specific women physicians; tapes and transcripts of oral history interviews with 41 women physicians in the U.S.; and materials on women as medical missionaries, women physicians in the armed services, and black women physicians. A photograph collection offers visual documentation of the lives and activities of women medical students and physicians, and the combined monograph and reprint collections consist of over 5,000 published items on women in medicine.

We have become aware over the past two years of several areas in which research is either lacking or incomplete. In the area of biographies and histories solid, definitive biographical publications have yet to be written on a number of women physicians who, while not Zakrzewska's, Preston's, Putnam-Jacobi's, or Blackwell's, are very interesting and important women. Further, the most recent comprehensive history of women in medicine was published in 1938, and while Dr. Kate Campbell Hurd-Mead, author of A History of Women in Medicine, accomplished a tremendous feat in writing this history, there are a few revisions which could be made and another 40 years to document. A historical profile of the Medical College of Pennsylvania, from its early years to the present, would prove very interesting and revealing, especially in terms of faculty and student attitudes, and type and quality of medical education and patient care as compared with other medical schools. Did a different type of woman chose an all-woman's medical school in the past? Today?*

Along the lines of a film project, a Women in Medicine film, suitable for career counseling, has not yet, to our knowledge, been produced. We also see a need for the development of an educational approach (film, books, etc.) designed to alter, at the pre-high school and high school levels, attitudes about women as primary health care providers.

Additionally, there are a number of surveys which might be conducted to gather current statistics (from women physicians) on the practice patterns of women physicians. A thorough survey of patient attitudes, taking into consideration many more variables than in the past, would give a better understanding of the political directions in which health care is going, as well as an idea of how the medical establishment can better meet the needs of the patient. A thorough survey of the current attitudes of women medical students and physicians would also be valuable and should take into account attitudes towards discrimination against women in medicine, political/feminist attitudes, and their approach or anticipated approach to patient care (the self-help movement, patient fees, etc.).

Another study which would prove very interesting and could provide a basis for looking more closely at (and perhaps altering) medical school curricula, would be a study of medical schools' approach in indoctrinating students in the matter of health-care-provider status, and all that that status involves. Are these attitudes consistent with feminist ideals? Will feminist ideals result in better patient care?

Finally, for a better understanding of how certain family, peer, and career counseling pressures affect a woman's choice to enter the field of medicine, a longitudinal study might be done of high school or junior high school girls and boys, following the group through pre-medical and medical training. Many of these projects, we feel, would be suitable for grant funding, while others might be considered for dissertation topics.

---

*Although the College began admitting males for the first time in its history in 1969, the composition of the student body is still 60% female.

SYMPOSIUM

MEDICAL PROBLEM SOLVING: A REANALYSIS AND ALTERNATIVE PERSPECTIVES

Chairman:	William C. McGaghie, Ph.D.
Participants:	William C. McGaghie, Ph.D. Anthony LaDuca, Ph.D. Eta Berner, Ed.D. John D. Engel, Ph.D.
Discussants:	Arthur S. Elstein, Ph.D. Sarah A. Sprafka, Ph.D.

393

SYMPOSIUM

Medical Problem Solving:
A Reanalysis and Alternative Perspectives

Organizer: William C. McGaghie, Ph.D.

Participants: Eta Berner, Ed.D.
John D. Engel, Ph.D.
Anthony LaDuca, Ph.D.
William C. McGaghie, Ph.D.

Discussants: Arthur S. Elstein, Ph.D.
Sarah A. Sprafka, Ph.D.

OBJECTIVES

The results of research reported in Medical problem solving: An analysis of clinical reasoning, authored by A.S. Elstein, L.S. Shulman, and S.A. Sprafka, have received widespread attention in medical education circles. Based on a psychological model of diagnostic inquiry, and through the use of sophisticated research procedures, the investigators' findings do not support the assumption that medical problem solving is a generic professional trait--even among physicians judged to be highly proficient diagnosticians. The purpose of this symposium is to consider the methodological procedures and findings reported by Elstein et al. from several perspectives and to suggest alternative paradigms for research on medical problem solving. The session has three primary objectives.

1. To critique the research reported by the authors of Medical problem solving, on both conceptual and methodological grounds.

2. To propose that features of the medical setting, and the content and structure of medical cases, receive increased attention when problem solving is studied and taught.

3. To provide a forum for discussion of medical problem solving from psychological, educational, and sociological perspectives.

Medical Problem Solving: A Reanalysis

William C. McGaghie, Ph.D.
Office of Medical Studies
University of North Carolina School of Medicine

Problem solving is a construct that receives much attention in the medical education community. Most medical curricula now emphasize skills involving acquisition and interpretation of patient data, formulation and testing of diagnostic hypotheses, and execution of treatment plans based upon carefully defined medical problems. Textbooks provide instruction on the construction, scoring, and interpretation of written simulations, i.e., PMPs, for purposes of measuring diagnostic and treatment skill (e.g., McGuire, Solomon, & Bashook, 1976). Extramural specialty boards also support the idea that the ability to solve medical problems underlies medical proficiency. For example, candidates for NBME certification submit to the Part III examination which purports " . . . to test aspects of clinical competence dealing with the ability to identify, to resolve, and to manage patient problems" (Hubbard, 1978, p. 40).

The widespread interest in medical problem solving--particularly diagnostic reasoning--has spawned research programs at several U.S. and Canadian universities (e.g., Johnson, Feltovich, Moeller, & Swanson, 1979; Barrows, 1977). One of the most recent, and certainly the most ambitious, studies on the subject has been published under the title, Medical problem solving: An analysis of clinical reasoning (Elstein, Shulman, & Sprafka, 1978). The book reports a series of investigations, done at Michigan State University, aimed at producing a clear description of the cognitive processes used by physicians to reach diagnostic decisions.

Principal among the research methods used by the investigators were three "high fidelity" simulations of physician-patient encounters. In contrast with less realistic exercises, a high fidelity medical simulation is one in which the behavior of a physician and a mock patient closely resembles that which occurs in a practice setting. The performance of physicians who served as research subjects was assessed for each simulation in light of a psychological model of diagnostic inquiry proposed by the investigators. The model involves four activities: (a) cue acquisition, (b) hypothesis generation, (c) cue interpretation, and (d) hypothesis evaluation. Fifteen dependent variables said to represent medical problem solving were identified. Elstein et al. report that the dependent variables "are essentially different combinations built on three basic building blocks: information search units, cues (and critical findings), and hypotheses" (p. 56).

Reported Results

A series of data analyses were undertaken to distinguish peer-nominated "criterial physicians" (highly proficient diagnosticians) from "noncriterial physicians" and to explain any observed differences in terms of the model of medical inquiry. Procedures including cluster and factor analysis were used to reduce the set of dependent variables and between group differences were sought using analysis of variance. In brief, the researchers were unable to demonstrate statistically significant group differences for any of the dependent variables. Among other findings, the data further suggest that diagnostic proficiency is highly case-specific, i.e., accurate diagnosis depends more on the content of a medical problem than on characteristics of medical persons.

Reanalysis

Selected portions of the researchers' data were reanalyzed to seek a deeper understanding of their findings. The correlation matricies (p. 87) involving the 12 continuous dependent variables for each simulation were factor analyzed using both conventional and nonmetric (Lingoes, 1973) techniques. In neither case did results from the reanalysis conform to those reported in the text. Whereas the authors allude to a three dimensional solution for the variables in each simulation, lending partial support to their proposed model, the reanalysis consistently produced a four dimensional solution. In addition, the pattern of variable loadings (including loadings for five of the variables selected by the researchers for detailed analysis) was inconsistent across the simulations. Plausible reasons for the conflicting results include (a) unreliability of the dependent variables; and (b) unstable correlations, and factor structures, due to the small number of physician/subjects.

These and other methodological weaknesses cast doubt not only on the findings of the high fidelity simulation studies reported by Elstein et al. but also upon the validity of their psychological model of medical inquiry. Subsequent presentations in this symposium give detailed treatment to the validity issue with particular reference to the content of the high fidelity simulations.

References

Barrows, H.S. (Organizer). Studies of the clinical reasoning process of medical students and physicians. RIME Conference Symposium, Washington, D.C., 1977.

Elstein, A.S., Shulman, L.S., & Sprafka, S.A. Medical problem solving: An analysis of clinical reasoning. Cambridge: Harvard University Press, 1978.

Hubbard, J.P. Measuring medical education (2nd ed.). Philadelphia: Lea & Febiger, 1978.

Johnson, P.E., Feltovich, P.J., Moller, J.H., & Swanson, D.B. Clinical expertise: Theory and data from the diagnosis of congenital heart disease. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

Lingoes, J.C. The Guttman-Lingoes nonmetric program series. Ann Arbor: Mathesis Press, 1973.

Solving the Medical Problem Solving Problem

Anthony LaDuca, Ph.D.
Director of Evaluations
Office of Medical Education & Evaluation
Hahnemann Medical College and Hospital
Philadelphia, Pennsylvania

> "There is a tendency to assume that although role is a 'purely' social matter, the engine that projects it - the person or individual - is somehow more than social, more real, more biological, deeper, more genuine. This lamentable bias should not be allowed to spoil our thinking."
>
> E. Goffman

In the pantheon of physician attributes none is more exalted that diagnostic problem solving. It is the stuff of legends, the touchstone of medical mystique. Not surprisingly, proficiency in diagnosis is prominent in the objectives of medical curricula and competence assessments, and medical educators and researchers have focused a great many of their efforts on this vital capacity. In particular the work of Elstein and his associates over the years has produced a singular contribution to investigations of clinical reasoning.

An impressive array of studies were conducted under the aegis of the Medical Inquiry Project (Elstein, Kagan, Shulman, Jason & Loupe, 1972; Elstein, Shulman & Sprafka, 1978a). One could not do justice to the Project in this format, and therefore, the following comments are restricted to those studies using physician subjects responding to "high fidelity simulations." used by "criterial and noncriterial physicians" in arriving at diagnoses. The investigators framed the problem of medical problem solving within a particular psychological research tradition, describing their efforts as "directed at development of an information-processing model" (Elstein, et al., 1978a, p.45). Their evaluation of their success in this endeavor is described elsewhere (Elstein, Sprafka & Shulman, 1978b). What is pertinent here is the findings from the high fidelity simulation studies.

Medical Problem Solving in Context

By their own admission the investigators' efforts were "largely unsuccessful," an assessment arising from two sets of results. First, it was found that the "criterial" physicians were not better diagnosticians than their "noncriterial" colleagues. In other words the distinction between these two study groups was invalid. Second, the data indicated "that intraindividual consistency on any of the variable are influenced much more by the structure of the problem and the individual clinician's understanding of its task demands than by consistency in individual problem-solving style" (Elstein, et al., 1978a, p.85). In less technical language this means that the performance of the physician subjects was inconsistent across the high fidelity simulations. As a result it was not possible to generalize or predict diagnostic proficiency across problems. In short medical problem solving was situation-specific.

These results cast doubt on the utility of the belief that medical problem solving is a cognitive factor (an introject, trait, disposition or innate ability). However, it should be pointed out that such a conclusion is consonant with some recent research and commentary (Bateson, 1972; Goffman, 1974; Hunt, 1975; May, 1977; Mischel, 1973). Frequently, these discussions have revived earlier arguments about the need to more carefully attend to the circumstances under which human behavior is studied (Hammond, 1948; LaDuca, 1979; Maher; 1978) and have proposed alternative educational research paradigms (Magoon, 1977; Shulman, 1970; Snow, 1974).

Medical Problem Solving and Medical Education

Diagnostic proficiency is of undeniable importance in the competent performance of the physician, and as result, proficiency in historytaking and physical examination is given high priority in the medical curriculum. At the same time, confidence in the problem-solving construct underlies a great deal of medical education practice, particularly in the evaluation of student or resident clinical performance.

1. Evaluation of senior medical students on clinical rotations raises the question of the equivalence, or comparability, of students' experiences in their separate rotations. The response has usually been that the patient problem content of the rotation is not as critical as the fact that the students get experience in the "process" of diagnostic problem solving. Elstein, et al., should give ample reason to reassess this position.

2. The use of oral examinations is still common in medical student evaluation and some specialty board certifications. Whatever the virtues of this method, the limited time restricts the number of cases considered. Thus, the same assumption of generic process is operating here as well, although it is not always articulated. This assumption is contradicted by the findings of the Medical Inquiry Project.

3. Evaluation of clinical competence during residency training and for specialty certification probably reflects more careful attention to the particulars of the clinical problems posed. However, Elstein, et al. report that their high fidelity simulations were "based on actual clinical records" and represented "problems that a general internist practicing in a community

hospital of moderate size could be reasonably expected to see..." (1978a, p.47). Despite this apparent content relevance the above mentioned situation-specific problem solving results were obtained. Thus, assurances that the problems presented to the resident are within the professional domain of relevance does not permit assumptions of generalizability of performance.

Avoidance of Lamentable Bias

The work of Elstein and his associates has produced results that fail to support belief in a general problem solving ability among physicians. These results are doubly significant.

First, the medical education community is confronted with evidence that refutes a basic and highly prized assumption. Tolerance of loose design in medical curricula, based on the belief that the particulars of clinical cases are not as important as the general diagnostic process, has been one consequence. Another is the questionable validity of assessment of diagnostic proficiency. Second, the conception of medical competence is disturbed. The ascendancy of situation-specificity at the expense of psychological introjects severely disrupts medical mystique.

Distasteful as it may seem, only one course of action seems indicated for future research in medical education. That is a direct attack on the issue of situation-specific performance in the form of careful analyses of the contribution of situation variables to the uniqueness of clinical problems. Such investigations should heed the calls for different research paradigms that permit analysis of the interaction between the physician and the circumstances posed by the clinical situation. This argument was beautifully summarized over thirty years ago.

> "...Proper sampling of situations and problems may in the end be more important than proper sampling of subjects, considering the fact that individuals are probably on the whole much more alike than are situations among one another."
>
> E. Brunswick

References

Bateson, G. Steps to an ecology of mind. New York: Ballantine Books, 1972.

Elstein, A.S., Kagan, N., Shulman, L.S., Jason, H. & Loupe, M.J. Methods and theory in the study of medical inquiry. Journal of Medical Education, 1972, 47, 85-92.

Elstein, A.S., Shulman, L.S. & Sprafka, S.A. Medical problem solving: An analysis of clinical reasoning. Cambridge, Mass: Harvard University Press, 1978 (a).

Elstein, A.S., Sprafka, S.A. & Shulman, L.S. The information processing approach to clinical reasoning: Reflections on the research method. Presented in symposium at Annual Meeting of American Educational Research Association, Toronto. April 1, 1978 (b).

Goffman, E. Frame analysis. New York: Harper Colophon, 1974.

Hammond, K.R. Subject and object sampling - a note. Psychological Bulletin, 1948, 45, 530-533.

Hunt, D.E. Person-environment interaction: A challenge found wanting before it was tried. Review of Educational Research, 1975, 45, 209-230.

LaDuca, A. The structure of competence in health professions. Presented in symposium at the Annual Meeting of the American Educational Research Association, San Francisco, April, 1979.

Magoon, A.J. Constructivist approaches in educational research. Review of Educational Research, 1977, 47, 651-693.

Maher, B.A. Stimulus sampling in clinical research: representative design reviewed. Journal of Consulting and Clinical Psychology, 1978, 46, 643-647.

May, R. Gregory Bateson and humanistic psychology. In Brockman, J. (Ed). About Bateson. New York: Dutton, 1977.

Mischel, W. Toward a cognitive social learning reconceptualization of personality. Psychological Review, 1973, 80, 252-283.

Mitchell, J.V., Jr. Education's challenge to psychology: The prediction of behavior from person-environment interactions. Review of Educational Research, 1969, 39, 695-721.

Shulman, L.S. Reconstruction of educational research. Review of Educational Research, 1970, 40, 371-396.

Snow, R.E. Representative and quasi-representative designs for research in teaching. Review of Educational Research, 1974, 44, 265-292.

Toward A Structure of Medical Problem Solving Situations

Eta Berner, Ed.d.

and

John D. Engel, Ph.D.

Center for Educational Development

University of Illinois, Medical Center Campus

Background

Numerous definitions of theory and the theory-building process in the behavioral sciences have been offered in recent years (Snow, 1973). Among the various conceptions of theory, Louis Guttman's view is most attractive, for it focuses on the connection between a conceptual framework and reality as represented by an empirical data structure. Guttman defines theory as "...an hypothesis of a correspondence between a definitional system for a universe of observation and an aspect of the empirical structure of those observations, together with a rationale for such an hypothesis" (Gratch, 1973). While Guttman's conceptions flow mainly from his interests in the field of sociology, parallels may be found in the field of psychology within Egon Brunswick's notions of probabilistic theory (1955). In fact, Brunswickian theory may be viewed as a particular instance of Guttman's notions of meta-theory.

Given Guttman's conception of theory, the conduct of inquiry becomes a search for matches between conceptualizations of data and structural aspects of data which evidence regularity. When these matches occur with sufficient regularity and generality, laws are derived. When matches are not found, three strategies are suggested: (a) modify the difinitional system for the universe of observations; (b) modify the structural aspect of the data to which the theory refers; (c) modify both definitional and data structures (Shye, 1978).

Behavioral science investigators are quick to employ procedures which possess quantitative elegance in their studies while disregarding the quality of the definitional base for making observations. Commenting on this phenomenon, Ebel (1962) has said "this quantitative sophistication of many ....is displayed not in the precision and elegance of their procedures for obtaining initial measurements, but rather in the statistical transformations, elaborations and analyses they are prepared to perform on almost any raw data given them." The result of this inappropriate behavior is, as was trumpeted by Brunswick in the early 1940's, an inability to achieve meaningful replications of empirical studies.

Shye (1978) notes that three lessors may be learned from the successful experience of the physical sciences:

> 1. ...the definitional system for collecting observations must specify the essential facets for which a partner is to be sought in the empirical data. This specification would permit meaningful replications and verifications and serve as a basis for corrections, additions and deletions.
>
> 2. ...concerning the aspect of the empirical structure of the observations, an open mind should be maintained about what kinds of such aspects could participate in a fruitful linkage with definitional systems. Adhering to an aspect of reality chosen in advance without weighing its potential for coupling with definitional systems may hinder or prevent the discovery of laws.
>
> 3. ...the process of discovering laws may be an arduous negotiation between the two parts of a theory, the definitional and the empirical, and therefore imagination and effort must be channeled toward the coupling as such, regarding each of the parts as a negotiator, willing to creatively modify, shift, remold, or turnabout its position in order to reach an agreement: (p. 4-5)

Given these lessons, Guttman's work relates to three areas.

First, the development of procedures for constructing and presenting definitional systems for a universe of observations. These methods are known as facet design and involve the construction of semantic constructions which map a universe of observations (i.e., mapping sentences).

Second, the analysis of alternative aspects of the empirical structure of observations to determine if they possess sufficient stability for the formulation of laws. These analyses include Smallest Space Analysis (SSA); Multidirectional Scalogram Analysis (MSA) and Partial Order Scalogram Analysis (POSA).

Third, the match between three areas constitute Guttman's facet metatheory which means the hypothesis that, "...in the long run, an optimal strategy for developing laws is the specification of certain formal roles for the facets in a mapping sentence. Together with the specific contents of these facets, these roles provide a rationale for an hypothesis of correspondence between the definitions of the mapping sentence and an aspect of the observed distribution of data.

An Alternative Approach to the Study of Medical Problem Solving

Recently, the most extensive and long term study of medical problem solving was concluded (Elstein, et.al., 1978). In summarizing their major findings, the authors comment:

> (d) Competence may be case related. Both physicians and medical students were found to vary considerably in their diagnostic effectiveness, according to the nature of the problem at hand. These findings are contrary to the widely held assumption that there is a clear (although not yet demonstrated) distinction between physicians who are "competent" and those who are not. Put another way, we may have to abandon our hope that competence, even in the subsidiary area of ability as a clinical problem solver, is a general characteristic. A more likely conclusion will probably be that the issue of physician competence (like the medical care process to which it applies) is more complex than we have wanted to acknowledge. Individuals will have to be evaluated by far more thorough procedures than is possible with a few formal examinations or clinical cases. We will need to think in terms of profiles of competence, in which individuals are regarded as being effective with particular types of problems, in particular situations (p. x)

We believe that this finding is a manifestation of a fundamental problem noted earlier, i.e., the social scientists' lack of formulating a reliable definitional basis for carrying out observations. Specifically, in the case of work on medical problem solving, there has been widespread neglect of sampling what Brunswick has referred to as representative instances of the "situation" (Brunswick, 1944, 1947, 1955).

The primary goals of our study are to define sufficient dimensions on which medical problems vary and to predict performance based on those definitions. Our work is being conducted in two separate but related phases. This paper will deal with the first phase of our work -- building a definitional system. A preliminary classification scheme for medical problem solving situations was built utilizing notions of Guttman's facet theory and Brunswick's probabilistic theory. Preliminary facets include dimensions such as: patient demographic characteristics; number of clinical problems; organ systems/disease types; treatment urgency; breadth of differential diagnosis; amount and type of data needed to arrive at diagnosis; complexity of management; initial and follow-up management strategies; and forms of treatment. These facets along with their associated facet elements yield a professional situational space containing over one million possible situations.

The remainder of our paper discusses the application of principles from facet theory to reduce this space. We conclude the paper with outlining a plan for programmatic research in this field.

References

Brunswick, E. Representative design and probabilistic theory in a functional psychology. Psychological Review, 1955, 62, 193-217.

Brunswick, E. *Systematic and representative design of psychological experiments*. Berkeley: University of California Press, 1947.

Ebel, R.L. Content standard test scores. *Educational and Psychological Measurement*, 1962, *22*, 282-288.

Elstein, A.S., Shulman, L.S., & Sprafka, S.A. *Medical problem-solving: An analysis of clinical reasoning*. Cambridge: Harvard University Press, 1978.

Gratch, H. (Ed.). *25 Years of Social Research in Israel*. Jerusalem: Jerusalem Academic Press, 1973.

Shye, S. On the search for laws in the behavioral sciences. In S. Shye (Ed.), *Theory construction and data analysis in the behavioral sciences*. San Francisco: Jossey-Bass, 1978.

Snow, R.E. Theory construction in educational research. In R.M.W. Travers (Ed.), *Second handbook of research on teaching*. Chicago: Rand McNally, 1973.

SYMPOSIUM

ALTERNATIVES TO TRADITIONAL METHODS OF CURRICULUM EVALUATION

Chairman: Paul L. Grover, Jr., Ph.D.

Participants:
Paul L. Grover, Jr., Ph.D.
Judith Sackoff, M.S.
Gerry R. Schermerhorn, M.A.
Michael M. Ravitch, Ph.D.
Ronald W. Richards, Ph.D.

Discussant: Henry B. Slotnick, Ph.D.

Alternatives to Traditional Methods of Curriculum Evaluation

Organizer: Paul L. Grover, Jr., Ph.D. - SUNY, Upstate Medical Center

In a review of a recent text on curriculum evaluation, Adrianne Bank cited a growing concern among evaluators that the empirically based, conventional prescriptions for curriculum evaluation have failed to satisfy the needs of evaluators, those who commission evaluations and those who are evaluated. She cited a need for "...a richness of descriptive evidence and a subtlety of interpretation about educational phenomena that is not satisfied by the designs and instruments currently used."* Based upon their experience as evaluators of courses, projects and curricula within their medical schools, the participants of this symposium share this concern. This symposium will explore non-traditional techniques which may be used to augment, amplify, verify and at times replace traditional methods of curriculum evaluation.

The objectives of this symposium are 1) to present a review of the limitations to which traditional methods of curriculum evaluation are subject, 2) to present a variety of alternative techniques with examples drawn from curricular and project evaluation, 3) to critically assess advantages and disadvantages of these alternative techniques, 4) to indicate how these alternatives can best relate to traditional methodology.

This symposium will consist of four prepared presentations and a critical analysis followed by open discussion. The first paper (Grover) will describe an application of the participant observer technique in gathering naturalistic data to evaluate a clinical clerkship. Examples of typical problem areas within the clerkship, the impact of such an evaluation, threats to objectivity and the costs of such an evaluation will be described.

The second paper (Sackoff), proposes a multiple-measurement approach to curriculum evaluation as a means of circumventing the interpretive problems introduced by reliance on a single measure. Qualitative data, content analysis, unconventional methods of data collection and non-reactive measures will be considered. A method for analysis of their content will be described.

A third paper (Schermerhorn), will describe three alternative techniques which are used to complement data gathered from student questionnaires and external measures. These techniques include responsive evaluation of a clinical clerkship, oral feedback sessions and content analysis of student comments.

The fourth paper (Ravitch and Richards) describes a multiperspective evaluation used in a self-study of the MSU Upper Pennisula Medical Education program. This method permitted the use of quantitative and non-quantitative data from diverse sources and resulted in the identification of specific program strengths and weaknesses.

In a prepared discussion, Dr. Slotnick will evaluate the studies presented from a consistent meta-evaluation perspective.

---

*Adrianne Bank, Center for Study of Evaluation, UCLA, in a review of the Handbook of Curriculum Evaluation by A. Lewy (Ed.) in AERJ, Vol. 16, No. 1, Winter 1979.

The Utility of Naturalistic Inquiry in Course Evaluation: A Case History

Paul L. Grover, Ph.D., SUNY - Upstate College of Medicine

A dissatisfaction with the forms, limitations and assumptions of traditional evaluation instruments and designs has led educators to consider alternative methods. Naturalistic inquiry has been proposed as a useful complementary technique and a viable alternative in certain situations. The purpose of this study was to evaluate a clinical clerkship in consideration of a proposed curricular change. A secondary purpose was to assess the utility of the naturalistic inquiry model as a means of gathering data on which to base such curricular decisions.

Traditional sources of evaluative data regarding courses in the medical school each present their own unique set of limitations. These limitations compromise their use as bases for making curricular decisions. Aspects of traditional techniques including student based questionnaires, peer and chairman review and student achievement data will be considered.

The naturalistic inquiry model will be contrasted with experimental and representative designs. Borrowing from anthropology and investigative journalism, theorists have advanced a model of inquiry termed, naturalistic (1). Where questionnaires, standardized testing and both experimental and representative designs impose their own preconceptions upon the situation under study, the naturalistic inquiry attempts to bring a "tabula rasa" to the situation. Its goal is to describe what is and look for patterns after describing. In this way, the biases of ommission and inclusion implied by these other techniques will be avoided.

The naturalistic inquiry model was applied in the evaluation of a clinical clerkship under review by the Clinical Curriculum Committee. The evaluation had the following objectives:

1. To accurately define and monitor changes in problem areas indicated by previous sources of evaluative data.
2. To gather data to inform and support decisions of the Clinical Curriculum Committee regarding the department's request for more teaching time.
3. To gather naturalistic data revealing previously unspecified problems and thereby contribute to a better understanding of the situation.
4. To obtain simultaneous verification of other evaluative instruments, specifically a revised course evaluation questionnaire.
5. To identify the advantages and disadvantages of the naturalistic mode of inquiry in course evaluation.

---

(1) Guba, Egon G., Toward a Methodology of Naturalistic Inquiry in Educational Evaluation, CSE Monograph Series in Evaluation No. 8, Center for the Study of Evaluation, University of California, Los Angeles, 1978.

The evaluator attended each of the experiences provided students in the clerkship including ward rounds, attending rounds, seminars, lectures, the operating room, clinics, course review sessions and the final examination. Those experiences which proved particularly informative were visited more frequently. Problem areas were defined and explored through a process of triangulation consisting of repeated observations and interviews with faculty, students and residents. The first two weeks of the six week period were devoted to a general survey and problem identification. The next two week period focused on each of these problem areas. The final two weeks were used to cross-check observations with those of students, faculty and residents.

Findings in each of ten focal areas identified will be mentioned briefly. Although related primarily to the course in question, they contain elements of general interest and may assist other evaluators in identifying key problem areas. Problems were identified in the clerkship organization, objectives, resident teaching attitude and performance, clinical teaching and clinical material. Strengths of the clerkship were identified in its lecture and seminar series, motivation of attendings, students' learning of tests and procedures and testing of students. The role of the course coordinator emerged as an important concern which was not previously identified. Recommendations were developed for each of these aspects of the course.

Advantages and disadvantages of the naturalistic inquiry model will be discussed. An evaluation conducted in this manner accumulates a large amount of concrete documentation. Such documentation proved useful in support of curricular decisions, staffing changes and changes in course content. The observer acquired an in-depth knowledge of the subject area assisting in the construction of behavioral objectives, evaluation techniques and other aspects of curricular design. In other words, the evaluator learned to speak the language of the specialty area and became a more helpful and credible change agent. This technique allowed accumulation of large amounts of evidence protecting positive aspects of the course from unnecessary changes. Finally, this approach revealed unanticipated factors which were necessary for course improvement but previously undetected by other evaluation techniques.

The naturalistic approach appears to have some disadvantages which require recognition and compensation. Principally, this approach requires a large time involvement. This author committed over 140 hours within six weeks to the project. Secondly, the author became aware of the danger of developing friendships with individuals, among the faculty, residents and students which could affect his objectivity. To prevent this, the naturalistic evaluator must make his role and the importance of his objectivity clear to all involved prior to beginning of the project. In a similar way, a personal commitment to a particular aspect of the course which is not well taught can lead one to overstress the importance of that area. On a technical level, drawing opinions from all of the participants and cross-checking with other sources of data will offset one's personal biases. On a personal and philosophical level, the best answer to these threats to one's objectivity is an awareness of their presence and their effect. The maintenance of a tentative, explorative and open attitude is the best response and in keeping with the spirit of this form of evaluation.

A Multiple-Measurement Approach to Curriculum Evaluation

Judith Sackoff, M.S., New Jersey Medical School

A multiple-measurement approach to curriculum evaluation is proposed as a means of circumventing the interpretive problems introduced by reliance on a single measure of a phenomenon. Since the mode of data collection contributes to the character of the data collected, analyses based on a single collection method necessarily incorporate the distortions and limitations of the method into the interpretation. Multiple-measurement depends on the simultaneous consideration of several measures collected by diverse methods to explore a single phenomenon. The validity of a given interpretation is thus corroborated by data collected by methods that, though individually biased, do not share the same biases. To adopt this methodology, curriculum evaluators must be willing to accept various data and methods of collection and analysis that they may have previously rejected. In particular, qualitative data, content analysis, unconventional methods of data collection and nonreactive measures must be considered.

In the present work, a multiple-measurement approach was used to determine the correct interpretation of student ratings of a basic-science course. The rating questionnaire consisted of a 21-item Likert-type scale covering various aspects of the teaching program and an open-ended question on the strengths and weaknesses of the course and individual faculty. Four additional sources of data on student opinion--all qualitative and collected by diverse methods--were used to corroborate the interpretation of the ratings. These included student responses to the open-ended question, a class letter, an editorial in the student newspaper, and notes of meetings between class representatives and the chairman. All were content analyzed. An interpretation of student opinion consonant with all of the data was derived. It was concluded that the student ratings alone provided neither an accurate nor a complete assessment of student concerns. By viewing the various data side-by-side and allowing for the strengths and biases of each, an interpretation emerged that differed significantly from the initial understanding based on the ratings alone, while nonetheless retaining elements of it.

More recently another curriculum evaluation problem has been encountered which the multiple-measurement methodology promises to illuminate. Initially, traditional methods of data collection and analysis were used to define the problem. The scores of NJMS graduates on the National Board Part III examination were consistently lower than on either Parts I or II, and inconsistent with other measures of clinical performance. It was hypothesized that scores on Part III, which tests the kinds of clinical skills that students develop in the fourth year of medical school and the first year of internship, were related to changes in the fourth-year curriculum; the correlation between Part III scores and the proportion of electives in the senior curriculum over an eleven-year period revealed that performance dropped as the proportion of elective time increased. To arrive at a full understanding of the basis for the observed relationship, the multiple-measurement approach has been applied. Methods, such as the following, that will illuminate the problem from diverse perspectives will be discussed: participant observation of electives and clerkships; a questionnaire survey of student opinion of electives; open-ended interviews with faculty involved with the elective program in present and past capacities; and documents tracing faculty and student perceptions of the purpose and effect of the program. The rationale, expectations and merits of such an approach will be contrasted to those of a more traditionally rigorous and inflexible one.

Complementary Approaches to Program Evaluation

Gerry R. Schermerhorn, M.A.
Southern Illinois University School of Medicine

A variety of methods for program evaluation has been used at Southern Illinois University School of Medicine. While evaluative data are collected by the more traditional means such as student questionnaires and external measures of program success, these methods are complemented by evaluation approaches capable of yielding additional and useful information.

In 1976, a responsive evaluation of the Medicine Clerkship was conducted. Based on the Stake model, a team of observers followed students as they progressed through the clerkship experience. Detailed time logs based on more than 250 hours of observation were analyzed and categorized resulting in a report that addressed such issues as students' use of time, interactions with faculty and residents, use of learning materials, patient mix and opportunities to perform clinical procedures. Variability in clerkship experiences based on the resident to whom a student was assigned as well as on the student's own level of assertiveness was investigated and documented. Frequent interviews with the observed students provided additional information regarding the students' perceptions of the clerkship.

When the final report was presented to the Department of Medicine later that year, their response could be characterized as interested but less than enthusiastic. A number of faculty thought the primary value of the study to be documentation of what they already believed. However, without claiming any direct causal relationship, I should like to point out that several problem areas defined and described by that responsive evaluation have since been addressed and in a couple of cases, resolved. Moreover, the Educational Resources unit that staffed the Medicine Clerkship study has since been asked to conduct similar studies of the Emergency Medicine rotation and the gastroenterology elective. While a particularly expensive mode of program evaluation, the responsive approach is useful when a comprehensive and highly detailed picture of a program is desired.

Another type of program evaluation is becoming a tradition at the SIU Springfield campus: oral feedback sessions. At the end of each major portion of the curriculum (referred to as a "sequence"), the class is divided into groups of approximately 12 students. Two members of Educational Resources meet with each group. One staff member acts as discussion leader; the other serves as recorder. A set of open-ended questions previously approved by the curriculum committee responsible for that sequence is used to stimulate discussion of a generic nature. Because students are generally given ample opportunity to evaluate each organ system block or clerkship through other means, they are asked to avoid comments specific to these smaller curriculum segments. Instead they are encouraged to look at the sequence experience as a whole, although comparisons of one block or clerkship with another naturally do occur.

The recorder notes whether the opinions voiced are those of an individual or seem to represent a consensus of the group. Furthermore, when the final report is developed from the individual report, it is noted whether criticisms or recommendations are representative of one or a number of groups.

Judging from the faculty response such reports have inspired, the information they provide is of interest and often results in faculty discussion and ultimately modification of portions of the curriculum. For example, in reviewing three years' worth of reports based on the feedback session following the clinical clerkship sequence, persistent themes exist, but a number of problems identified early are notably absent in subsequent reports. As an additional measure of this method's success, similar oral feedback sessions are now regularly conducted at the end of several of the clerkships.

One of the most obvious advantages of the oral sessions is that they potentially offer more data, since people are usually inclined to say more than they would write on an evaluation form. Additionally, the interactive quality these sessions assume results in one student's comment inspiring a supportive or contradictory one that otherwise might not have been expressed. Issues arise that were not foreseen by the evaluator and therefore would not have appeared on a standard questionnaire.

A third type of program evaluation is used exclusively in the first and second sequences of the curriculum. Comment pads are placed outside lecture halls, and students are encouraged to vent their opinions about lectures, labs, clinical experiences, learning materials, examinations and even school facilities. This is another interactive approach, and one student's comment is often challenged or endorsed by a subsequent contributor. Most importantly, the comments are reviewed frequently and sometimes result in timely modifications. When students realize their opinions count, they continue to provide useful recommendations and constructive criticism.

These three methods of program evaluation were never intended as substitutes for the traditional, more readily quantitative approaches. Rather, we have found that the relatively "soft" data they provide are useful in determining program successes as well as areas needing improvement.

Multiperspective Evaluation of a Rural Undergraduate Medical Education Program

Michael M. Ravitch, Ph.D., Ronald W. Richards, Ph.D., Office of Medical Education
Research and Development, Michigan State University

The evaluation of the Upper Penninsula Medical Education Program (UPMEP) was motivated by the need to provide formative information to UPMEP administration and faculty at the end of the program's first full (four year) cycle, and at the same time to provide summative information to the Liasion Committee on Medical Education as part of the general reaccrediation of the Michigan State University College of Human Medicine (CHM). The major challenge of this evaluation was to incorporate many kinds of data from many sources, and to reconcile the comments and opinions of individuals who had had UPMEP experience ranging from limited to extensive and from specific aspects of the program to the whole program.

The UPMEP is a primary care track established by CHM in cooperation with the Upper Penninsula Health Education Corporation in 1974 to provide undergraduate medical education in a rural context and as part of a larger plan to improve the quality of health care in the Upper Penninsula.(1) Ten students, who applied jointly to CHM and UPMEP, are admitted in alternate years. After an orientation term (Phase I) in East Lansing, the students spend the next three and one-half years in the Upper Penninsula, with some elective experiences elsewhere. The UPMEP follows the CHM Track II small-group problem-based basic science curriculum (Focal Problems) in years 1 and 2.(2,3) Based in a family health center (FHC) the UPMEP affords students early longitudinal contact with patient care. Some contacts in the FHC and elsewhere correspond to the clinical cases in the Focal Problems series. In the third and fourth years (Phase III), students receive clinical instruction in a 36-week ambulatory-based Comprehensive Care Clerkship (CCC), comprising medicine, pediatrics, surgery, obstetrics and gynecology, psychiatry, family practice, and a behavioral science seminar. A 12-week Hospital Care Clerkship (HCC), (henceforth 24-weeks), an indepth experience in the care of hospitalized seriously ill patients; and electives are also part of Phase III.

Method: The evaluation was guided in its attention by a list of goals and areas of interest identified by the Director of UPMEP, but for this symposium we will restrict our attention to the area of student performance. Given this remote (400 miles from East Lansing), innovative program, substantial attention was given to assessing the quality of preparation, evidenced in performance, of the students CHM would graduate from UPMEP.

For student performance, as for other goals or areas of interest, multiperspective methods were used (1) to identify available or needed data which would be responsive to the evaluation objectives and (2) to identify the perspective associated with each data source.(4) Perspective dimensions included the duration of a person's experience (one month to four years), the extent of contact (occasional interaction, regular interaction), program area (Phase or course or clerkship), and student contact (1-20). For a few sources, strong existing biases, supportive or critical, were often noted. Data were literally tagged with indicators so that comments, ratings, and scores could be placed in perspective in reaching conclusions. The multiperspective approach involves identifying the limitations of data from different perspectives (cf. triangulation).(5) In order to estimate the reliability and validity (generalizability)(6)

412

of the results.

Data: Phase II data included multiple-choice examinations using cumulative (across focal problems) scores in each basic science (anatomy, biochemistry, etc.), as well as scores on problem-solving exercises. Phase III data included CHM departmental examinations, preceptor ratings and comments, evaluation by CCC directors, HCC directors, and elective preceptors. Additional information was obtained from questionnaires prepared for this evaluation and distributed to UPMEP faculty, students, elective preceptors, CHM departmental chairmen, and CHM faculty involved with the UPMEP. Information was also obtained from students' files and from minutes of regular UPMEP faculty meetings. Further information on the clinical preparation of 4th-year students was obtained from a report by a team of six physicians from other medical schools, who were invited by the Director of UPMEP to observe and interview UPMEP students.

Results: In Phase II, UPMEP students performed as well as other CHM students on all measures and across basic science areas. In Phase III, UPMEP students passed all clerkship requirements, including examinations, preceptor ratings, case writeups, and general evaluations. Multiple-perspective analysis identified themes which were consistent across sources of comments and ratings: UPMEP students were considered particularly strong in clinical interview and examination skills, relating to patients, ability to deal with sensitive problems, independent learning, responsibility, and intellectual honesty. Regarding knowledge and clinical skills, students generally performed well on examinations and in clerkships. Clinical faculty, the team of visiting clinicians, and elective preceptors all indicated that students are well prepared to be competent clinicians. No problems were noted which would cause major concern about their performance.

Conclusions: The evaluation concluded that UPMEP students were generally well prepared, with particular strengths in several important areas related to clinical practice. Recommendations following from the evaluation include: continuing the ambulatory care experience, maintaining program aspects promoting independent learning, maintaining the experience in longitudinal care of patients, continuing the problem-solving approach to the study of clinical problems, and expanding the recently implemented Phase III HCC seminar in basic sciences related to current patient problems.

The multi-perspective approach was difficult, but appropriate. Tagging data by perspective dimensions was tedious, but subsequently it was easy to identify salient data (for a particular objective) and to assess the limitations of their validity. The greatest work was in the initial categorization of perspectives, and in the subsequent sorting of sources. In an innovative program, strong opinions are readily available; the greater the departure of the program from the conventional, the greater the polarity of attitudes pro and con, and thus the greater the challenge for the evaluator. The multi-perspective approach permits some sorting out of the data, according to validity, limitations of experience (e.g. time and program, part of curriculum, number of students observed) and of attitudes. The results are conclusions, balanced and broadly based, supported by multiple sets of data. In some cases, no balance occurs, and this result can be used to recommend improvements in the data base for program monitoring and evaluation.

The press of the LCME review tended to skew the evaluation toward examining the issue of how UPMEP graduates compare with other medical school graduates. Subsequent studies will be tailored toward examining students' achievement of UPMEP goals.

Addenda: All ten graduates passed the FLEX examination. In a follow-up survey of UPMEP graduates and their residency directors, graduates were rated well prepared for their clinical responsibilities, both compared to demands of the job and compared with the preparation of graduates from other medical schools.

References

1. Richards, Ronald W., Werner, Paul T., Weston, W. Donald, and Ways, Peter O. "The Upper Penninsula Medical Education Program: Educating Primary Care Physicians for Rural Areas." presented at the Kellogg Conference, Medical Education Since 1960-Marching to a Different Drummer, Michigan State University, 1978.

2. Echt, Robert and Chan, Sui-Wah. "A New Problem-Oriented and Student-Centered Curriculum at Michigan State University." Journal of Medical Education, 1977, 52:681-683.

3. Ways, Peter O., Loftus, G. and Jones, J.M. "Focal Problem Teaching in Medical Education." Journal of Medical Education, 1973, 48:565-571

4. Ravitch, Michael M. "Multi-Perspective Approach to Evaluation." Newsletter-Research on Evaluation Program, 1978, 1:9.

5. Campbell, Donald T. "Distinguishing Differences in Perception from Failures of Communication in Cross-Cultural Studies." in F.S.C. Northrop and H.H. Livingston (Eds.), Cross-Cultural Understanding: Epistemology in Anthropology. New York: Harper and Roe, 1964.

6. Cronbach, Lee J., Gleser, Goldine C., Nanda, Harinder, and Rajaratnam, Nageswari. The Dependability of Behavioral Measurements. New York: John Wiley and Sons, 1972.

SYMPOSIUM

NEW HORIZONS IN COST CONTAINMENT CURRICULA

Chairman:	Mohan L. Garg, Sc.D.
Participants:	Michael J. Garland, Ph.D. William B. Applegate, M.D. John M. Eisenberg, M.D.
Discussants:	John G. Freymann, M.D. James I. Hudson, M.D.

SYMPOSIUM

NEW HORIZONS IN COST CONTAINMENT CURRICULA

Organizer: Mohan L. Garg, Sc.D.
Director, Dept. of Cost Containment & Evaluation
Professor, Dept. of Medicine
Medical College of Ohio at Toledo
C.S. # 10008, Toledo, Ohio 43699

The purpose of this symposium is to present several recent, innovative curricular activities in the area of cost containment, including:

1) The expansion of curricular efforts from the undergraduate to the graduate levels of medical education.
2) The incorporation of several ethical dimensions as they directly affect cost containment objectives.

As a consequence of the initial success of curriculum development efforts at the undergraduate medical education level, new attempts are being made in selected medical schools to extend these program efforts to the graduate level. These new programs will be described by their developers and will highlight ethical considerations that have not yet been explicitly addressed by other programs. Critics of cost containment programs often claim that the patient's optimal well being becomes a victim of benign neglect. These new programs incorporate such concerns into the training process building on a framework that posits values as inextricable components of any cost-effective decision-making process.

Two presentations will focus on graduate medical programs, with some empirical evidence of the relative success achieved through such efforts. The other presentation will describe a unique attempt to systematically highlight the ethical dimensions in the decision making process and their integration with the general objectives of cost containment programs.

The discussants bring unique, well-known, and relevant experiences to this subject that will effectively synthesize the issues raised by the session.

ABSTRACTS

ETHICS AND ECONOMIES:
THE CHALLENGE OF COST CONTAINMENT EDUCATION IN THE MEDICAL SCHOOL

Michael J. Garland, Ph.D.
Dept. of Public Health & Preventive Medicine
University Oregon Health Sciences Centre

Voices from many sectors of society have begun to form a chorus calling for containment of medical care costs. Among the means being advocated is that of educating providers about the actual costs. That educational effort should, it is urged, include those preparing to enter the field.

From the educator's standpoint two major questions arise in response to urgings to include cost containment education in the medical school curriculum: "Where?" and "When?" The curriculum is by common agreement already overcrowded. Is there room for yet another required course? If not required, will an elective reach enough students to be effective? As to "When?" should the course be presented during the pre-clinical or the clinical years? The approach described here seeks to integrate some cost containment education serially throughout the four year curriculum using parts of various courses to present the total program.

The content of such a program has to be carefully looked at. Two assumptions must be examined: (1) that there is something medical students can learn that will lead them to be cost effective allocators of medical resources; (2) that medical students can become cost effective allocators of medical resources without (a) sacrificing quality in medical care, (b) violating their duties to patients, (c) impeding progress in medical science, or (d) increasing inequities in the medical care system.

First assumption. The knowledge about cost containment is intended to issue into action, namely the efficient allocation of medical care resources. The path from knowledge to action is brought about under the influence of values and my means of complex judgments. The knowledge base includes; (1) descriptive empirical data concerning U.S medical care in both particular and aggregate form; (2) descriptions of currently employed reimbursement mechanisms; (3) identification of economic incentives built into the present system; (4) comparisons among various health care systems; and (5) current research goals and problems related to progress in medicine. This material can be learned as well as any empirical information.

More than empiric information is required. Values are inextricably involved. Five categories of values can be identified for analyzing the transition from knowledge to action in the provision of cost effective medical care: (1) rational self-interest, (2) respect for persons, (3) rights, (4) justice, and (5) the common good.

In existence, knowledge and value issue into action via the exercise of judgment whereby one engages one's own freedom to the task of enacting value in the world of physical and personal relations.

The interplay of knowledge, values and judgment is well displayed in the educational technique of the case method. In the case empirical knowledge and values must be brought to bear on a problem (or set of problems) to be solved. The resolution chosen is revelatory of the possession of relevant empirical knowledge and the relative weight of personal values. The case method mimics social existence while recognizing that existential ethical judgments require decisions from which consequences follow relentlessly and resist easy alteration.

Second assumption. If cost effectiveness in medical care is to be taught, it must be developed in full awareness of certain legal, ethical, and social constraints on acceptable methods of containing costs. Teaching by the case method is well suited to pursuing particular goals (eg., cost effective medical care) under the constraints of other societal goals and values (eg., quality medical care, professional duties, pursuit of progress, and reduction of inequities). Empiric knowledge, value awareness, and judgmental acuity are all required in order to respond to the second assumption.

Conclusion: Cost containment medical education is a gamble that something can be taught to and learned by medical students in the course of their training that will influence their later behavior as medical care providers. Medical students represent only one of several groups all of whom must behave in concert if health care cost inflation is to be brought under control. Cost containment educational efforts in medical schools can at best be a contribution towards a solution and their ultimate effectiveness is dependent on actions taken within other groups. Still these curricular efforts are important because of the central role physicians play in the actual allocation of health care resources.

COST CONTAINMENT PROJECT UNIVERSITY OF NEW MEXICO

William B. Applegate, M.D.

The Departments of Medicine, Family Medicine and Pediatrics at the University of New Mexico are currently conducting a cost containment project in their ambulatory clinics. Only limited data is available from Family Medicine and Pediatrics, so this report will only contain data from the Internal Medicine portion of the project.

During the first year of the project no intervention was attempted. Baseline data was obtained on the charges resulting from housestaff encounters with patients in the adult medicine clinics. Charge data has been extracted directly from the medical record for all housestaff encounters with patients who carry the diagnoses of hypertension, adult onset diabetes, congestive heart failure, degenerative joint disease or bronchitis. Patients with combinations of these diagnoses were included, and each diagnosis was assigned a seventy rating based on pre-set criteria. In addition, the number of other problems listed on the chart was also coded. The charges coded include number of visits, lab tests ordered, drugs prescribed, consultations obtained and admissions. Data will be presented at the meeting regarding the effect level of resident training and patient payment status have on average charge per encounter and per patient over a six month period. Additionally, data will be presented on the range of charges by diagnosis.

During the first year a medical record audit has been carried out on a sample of each resident's patients. The audit is a process audit involving only patients with the specific diagnoses mentioned earlier. Criteria for the adult were established by a panel of physicians which included a mix of academic and private practice physicians. The results of the audit will not be part of the feedback. The process audit is being used as a measure of quality to answer the question: if average costs decline during the intervention period, does this have any impact on the process audit?

We are currently in the second year of the project which is the intervention year. The intervention consists of the following: (1) Every quarter each housestaff member receives feedback on his average charge per encounter by diagnosis. This is compared with the mean and range for all medical residents. (2) Three didactic sessions have been conducted by the faculty with all housestaff in the clinic. The sessions have dealt with pertinent aspects of medical economics and decision theory. (3) Four patient management ambulatory case simulations have been developed which each resident works through and then discusses his/her choices with a designated attending. (4) Each month the residents receive sample itemized bills from selected patients in their practice. (5) Clinic attendings have been required to review at least two cases per month per resident to conduct informal discussions of the cost-effectiveness of the decisions made.

Since we have a before/after study design, comparison will be made between residents at each level in hears one and two of the study rather than between individuals over a period of time. Data from the first quarter of the intervention year will be available at the time of the presentation.

AN EDUCATIONAL PROGRAM FOR REDUCING THE OVERUTILIZATION OF DIAGNOSTIC PROCEDURES

John M. Eisenberg, M.D.
Sankey V. Williams, M.D., Linda A. Pascale, M.S.
Hospital of the University of Pennsylvania

Because of their role in generating health care expenditures, physicians have become a prime focus in recent efforts to control costs. Particular attention has been paid to the physician in training whose cost-generating habits have not yet been formed. To encourage voluntary cost containment, medical educators have introduced a variety of training methods into undergraduate and graduate medical programs (1).

While few have been able to demonstrate a direct connection between educational intervention and changes in physicians' utilization of services, some who have been successful have reported a combination of educational and administrative interventions which may have contributed to the change (2,3).

In planning a cost containment program at the Hospital of the University of Pennsylvania, our first objective was to identify the causes for overutilization at our hospital. Since housestaff order all inpatient laboratory tests on our medical service, we surveyed them to determine the reasons for excessive testing. Factors which they considered to be most important were those related to inexperience and poorly-developed habits--factors which would seem to be responsive to education.

We introduced an intensive educational program in the use of diagnostic tests. Our objectives were:

1. to provide physicians with the technical information necessary to make appropriate decisions concerning diagnostic services utilization
2. to improve physicians' awareness of the financial consequences of medical decisions
3. to improve physicians' awareness of their own cost-generating behavior.

Housestaff Program

At an introductory grand rounds, we traced the history and the economics of the increase in health care expenditures, concentrating primarily on diagnostic services utilization. In addition, we reviewed the literature on the value of multiphasic screening.

We have used monthly medical management conferences to address more specific issues. At our first conference, we presented a videotaped patient simulation. The simulation deals with the diagnosis and treatment of a patient with a renal stone. The videotape contrasts the diagnostic techniques of an experienced physician with those of one who is still in training. The point is stressed that, due to his expertise, the experienced physician requires fewer and less expensive procedures in making a diagnosis. While physicians in training might require more information (more procedures) in making a diagnosis, the goal should be to provide quality care with a minimum of risk and cost to the patient.

A patient management exercise was distributed prior to the showing of the tape and was used to set the scene for the viewers. The exercise lists procedures and services that a physician might order in treating a patient who presents in the emergency room with abdominal pain. By completing the exercise, viewers were able to compare their own ordering patterns with those of the physicians on the tape.

A newsletter concerning diagnostic services utilization has been developed and is distributed monthly to medical faculty and housestaff. Newsletters are designed to educate physicians about the diagnostic value and limitations of specific procedures relative to the fees charged for them. In a continuing effort to improve physician awareness of the costs of health care, we have devoted two issues of the newsletter to publication of the charges for more frequently used diagnostic tests. Individuals from a variety of Hospital and University departments have demonstrated their support of our objectives by writing or editing issues of the newsletter. For instance, the chief resident in pathology wrote an issue which deals with appropriate use of protein determinations; the vice chairman of the Department of Radiology authored an issue on the use of the portable x-ray examination; and the director of the hypertension clinic wrote about the diagnostic evaluation of the patient with hypertension.

In a more comprehensive version of a pilot program (4), we have instituted a program of feedback to housestaff. With the help of a panel of twelve physicians from the Department of Medicine at the University of Pennsylvania, we developed criteria for appropriate repetitive use of ten diagnostic tests. Cases are identified for review by a computer-based system designed to detect repetitive use of inpatient laboratory tests. The computer selects for review those cases in which multiple determinations take place within a specified time frame. Criteria are applied to these selected cases by medical records technicians who conduct daily chart reviews. Through this method, physicians who order excessive procedures are identified. These physicians are notified of instances of unnecessary testing through letters placed on their patients' medical charts.

Medical Student Program

Tied into the housestaff program is a learning unit for medical students. The unit is presented as part of the clinical curriculum in medicine. Students participate in a four-week program which includes several lectures, readings and seminars on cost containment. The didactic materials used in the housestaff program have been adapted for the student program. For instance, newsletters are selected for discussion to coincide with topics in the clinical curriculum. Students also participate in cost-utility discussions related to the care provided to current inpatients.

In addition to direct classroom training, students are provided with housestaff role models who have been sensitized to cost-containment issues through the program outlined above.

References

1. Friedman, E. Changing the course of things: costs enter medical education. Hospitals, J.A.H.A. 53:82-85, 1979.
2. Griner, P. F., et al. Use of laboratory tests in a teaching hospital: long-term trends. Annals of Internal Medicine 90:243-248, 1979.
3. Roark, A. C. With medical bills soaring, can young physicians learn to be cost-conscious? Chronicle of Higher Education 17:7-8, 1978.
4. Eisenberg, J.M., Williams, S. V., et. al. Computer-based audit to detect and correct overutilization of laboratory tests. Medical Care 15:915-921, 1978.

SYMPOSIUM

PATIENT/HEALTH EDUCATION: TRAINING FOR WHAT?

Chairman:	Gary M. Arsham, M.D., Ph.D.
Participants:	Gary M. Arsham, M.D., Ph.D. Edward E. Bartlett, M.P.H. Ellen J. Cohen, Ph.D. Wendy Diane Squyres, Ph.D.
Discussant:	Merlin K. DuVal, M.D.

SYMPOSIUM

PATIENT/HEALTH EDUCATION: TRAINING FOR WHAT?

Organizer:	Gary M. Arsham, M.D., Ph.D. Pacific Medical Center San Francisco
Participants:	Gary M. Arsham, M.D., Ph.D. Edward E. Bartlett, M.P.H. Ellen J. Cohen, Ph.D. Wendy D. Squyres, Ph.D.
Discussant:	Merlin K. DuVal, M.D.

Objective:

Patient and health education are receiving increasing attention from clinicians, educators, legislators, and bureaucrats because of the potential for reducing costs and improving the quality of medical care. However, not everyone agrees with the evidence supporting this position. This symposium will present two complementary frameworks for training health professionals to do patient education, and explore the value of patient and health education from two opposite perspectives. Attendees will be better able to judge the worth of implementing such training programs in their own institutions.

Symposium Introduction - An Algorithm for Patient Education*

Gary M. Arsham, M.D., Ph.D.
Pacific Medical Center
San Francisco

This algorithm (see figure) illustrates a simple approach to patient education that stresses:

1. Concurrence between patient and provider on goals and desired outcomes, and

2. Behavioral diagnosis of the causes of inappropriate health behavior.

Although most algorithms are intended to be rather rigorously followed, this one is intended primarily to stimulate the practitioner to consider a variety of factors in patient education. The algorithm is organized into a cycle that includes three familiar categories (see bottom of the algorithm): 1. educational (behavioral) diagnosis, 2. educational treatment, and 3. monitoring and assessment. The competencies that comprise these categories are beginning to be identified and made more explicit, as will be described in the presentation by Ed Bartlett, to follow.

The first category, educational diagnosis, includes discussion between patient and provider over goals and outcomes, as well as an analysis of the causes of health behavior deficiencies. Most patient education is intended to result in behavior change; therefore, it is vital that the patient agree with the desired outcomes. Although there may be a variety of ways of achieving a given outcome, it is the patient who must take major responsibility for the necessary changes. Ellen Cohen will discuss the relative power of the individual to control important components affecting health behavior, and review the health and patient education literature in this context. I believe that, with "educational" help, patients can learn to exercise more control over the environment and their responses to it.

The next step, after identifying and discussing the health behavior deficiencies and agreeing on desired outcomes, is to analyze the causes of the deficiency with the patient. The deficiencies can be organized into four categories:

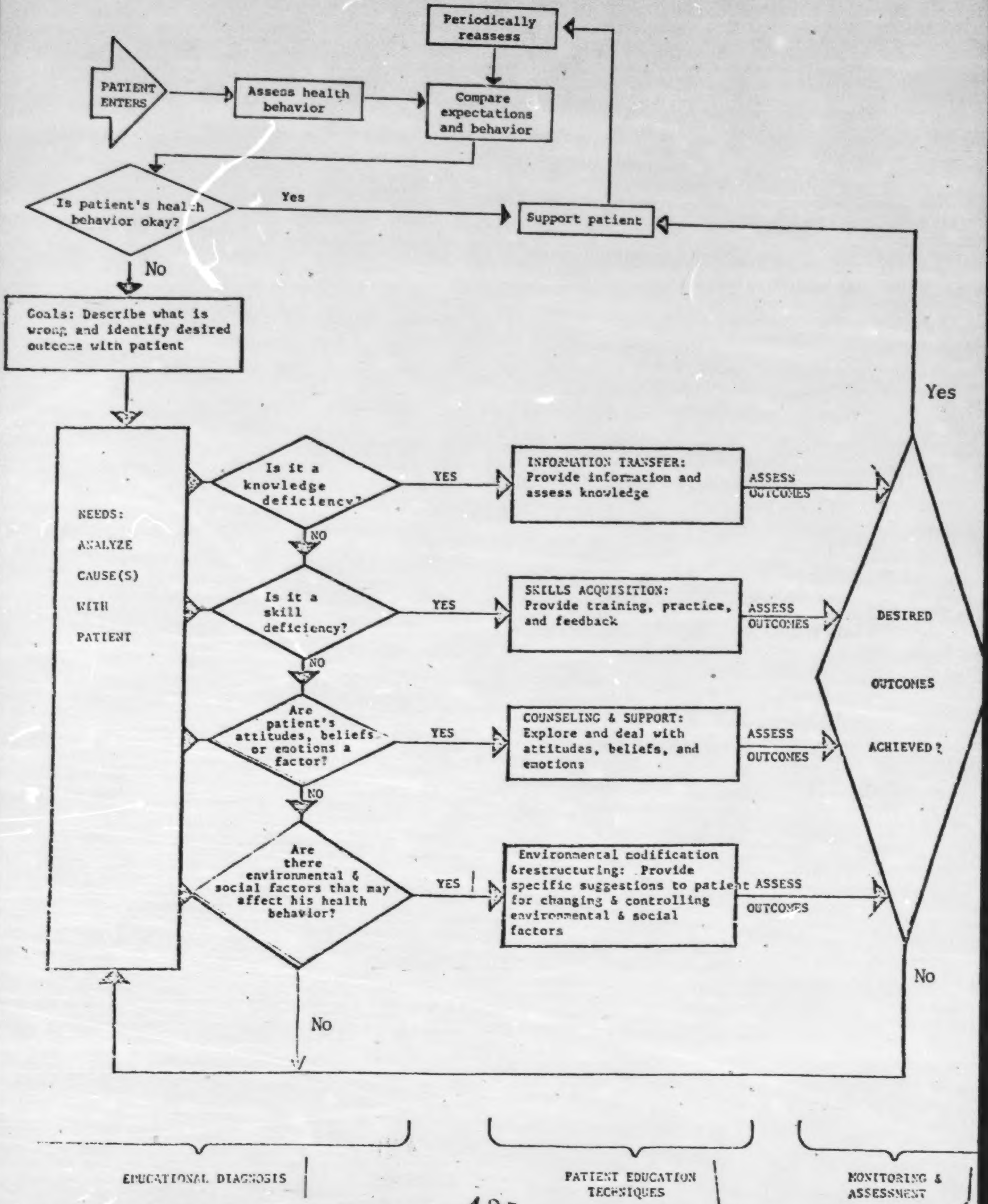
1. Knowledge deficiency
2. Skill deficiency
3. Attitudinal, belief, or emotional barriers
4. Environmental or social barriers.

In most patients the causes are attributable to a combination of these

---

*This approach is the result of ideas derived from many sources whose help is gratefully acknowledged. Some sources include: Larry Hulbert, Gary Kahn, August Colenbrander, Donald Bosshart and the TOPPE Task Force on Training of Physicians in Primary Care, and especially Robert Mager and Peter Pipe.

FIGURE: STEPS IN PATIENT EDUCATION


PATIENT ENTERS
Assess health behavior
Compare expectations and behavior
Periodically reassess
Is patient's health behavior okay?
Yes
Support patient
No
Goals: Describe what is wrong and identify desired outcome with patient
NEEDS: ANALYZE CAUSE(S) WITH PATIENT
Is it a knowledge deficiency?
YES
INFORMATION TRANSFER: Provide information and assess knowledge
ASSESS OUTCOMES
NO
Is it a skill deficiency?
YES
SKILLS ACQUISITION: Provide training, practice, and feedback
ASSESS OUTCOMES
NO
Are patient's attitudes, beliefs or emotions a factor?
YES
COUNSELING & SUPPORT: Explore and deal with attitudes, beliefs, and emotions
ASSESS OUTCOMES
NO
Are there environmental & social factors that may affect his health behavior?
YES
Environmental modification &restructuring: Provide specific suggestions to patient for changing & controlling environmental & social factors
ASSESS OUTCOMES
No
DESIRED OUTCOMES ACHIEVED?
Yes
No
EDUCATIONAL DIAGNOSIS
PATIENT EDUCATION TECHNIQUES
MONITORING & ASSESSMENT

factors. In the past, patient education has focused predominantly on correcting knowledge deficiencies, often when such a deficiency may have been only a minor factor in the patient's problem. More recently, increased attention has been paid to the personal, social, and environmental factors that can affect health behavior. In my definition of patient education, I include any patient-provider interaction that can help the patient deal with such factors. Wendy Squyres will analyze various definitionsof patient and health education, as well as review the evidence supporting the utility of patient and health education. Merlin DuVal, as discussant,will help us criticize and synthesize the ideas presented in the symposium.

The algorithm next leads to the educational treatment or treatments, which follow directly from the causes that have been identified. Finally, the cycle is completed as the originally identified outcomes are monitored on recurrent visits. If they have not been achieved, the cycle is repeated with a fresh analysis.

This algorithm is a useful tool for introducing a multi-factorial approach to patient education in a relatively simple way. The four causal categories provide an easy conceptual handle for most practitioners to grasp, increasing the likelihood that the educational treatment will be based on an analysis of multiple factors, rather than one or two.

Identifying the Desired Patient Education Competencies of Primary Care Physicians: Results of a Delphi Survey

Edward E. Bartlett
Division of Health Education
Johns Hopkins University

Introduction

Education of the patient long has been recognized as an important part of high quality medical care. The increasing awareness of the problem of non-adherence to the therapeutic regimen, and the emphasis upon lifestyle modification as a preventive measure, have led to a greater understanding of the role of patient education in medical care. Traditionally, the processes of patient education have been taught to medical students and residents primarily through role modeling. However, with recent advances in the conceptual, theoretical, and empirical basis of patient education practice, it appears that a more systematic approach to teaching behavioral diagnosis, educational prescription, and other health education concepts is needed. Thus, a growing number of medical schools and primary care residencies have introduced formal didactic instruction in health education into their curriculum.

The best educational curriculum is one based upon specific observable competencies (skills) which serve as the basis for the development of educational content, teaching methods, and evaluation techniques (1). To this end, this project attempted to identify the desired patient education competencies of primary care physicians. Additionally, 15 medical education programs (predominantly family practice residences) volunteered to collaborate in the project to implement the identified competencies.

Methodology

The Delphi survey technique, first developed at the Rand Corporation to attempt to forecast future events (2), was selected to achieve consensual validity regarding the desired competencies. Because a broad range of opinions was needed, approximately equal numbers of persons from six groups were selected: practicing primary care physicians, medical educators, health education specialists, nurses, health consumers, and other health professionals (pharmacists and health manpower planners). A nonprobability sample from these groups were selected because of the difficulties in specifying the universe for these groups, and because of the iterative, interactive nature of a Delpi survey (3).

The Delphi Survey was accomplished in two Rounds. In Round 1, the participants suggested any competencies, and their rationales, which they thought were relevant for one of three tracer conditions: essential hypertension, routine prenatal visit, and pediatric otitis media. In Round 2, the panel reviewed all the competencies listed in Round 1, and indicated whether they thought they were usually necessary, sometimes necessary, or seldom necessary. Response rates to Round 1 and Round 2 were 50% (45 persons) and 72% (42 persons), respectively. Substantial editing was required after both Rounds to improve syntax and to eliminate redundant competencies.

Results and Discussion

All of the competencies were organized using a problem-solving framework analogous to the Problem-Oriented Medical Record (4). 88 competencies were identified by respondents to the three tracer conditions, of which 70 were rated by 50% or more of the panel as usually necessary. Only four of the competencies were rated by 50% or more of the participants as seldom necessary. Examples of some of the competencies, and their ratings, were:

Competency Number	Competency	% Usually Necessary	% Sometimes Necessary	% Seldom Necessary
10	Assess patient's priorities for behavior change.	88	13	0
16	Assesses patient's sense of confidence of ability to manage condition.	50	50	0
32	Uses behavior modification techniques, as appropriate, e.g., contracting for weight loss.	26	74	0
48	Demonstrates tolerance of disagreement and/or non-adherence - shows patience.	84	16	0
68	Checks patient's kept appointment rate.	67	33	0
74	Documents patient education as a plan element in the POMR .	74	25	2

It is probable that the usually necessary competencies will not be used in every patient encounter, but nonetheless will be useful to the primary care physician for a variety of conditions. Additionally, the list of recommended competencies must be adopted to the exigencies of the local situation. Currently, a modular curriculum based on these competencies is being developed. Parts of this curriculum already have been pilot-tested, with generally favorable response from faculty, residents, and others.

References

1. Segall, AJ, et al, Systematic Course Design for the Health Fields, New York, John Wiley and Sons, Inc., 1975.
2. Linstone, HA, and Turnoff, M (eds.), The Delphi Method: Techniques and Applications, Reading, MA, Addison-Wesley Publishing Co., 1975.
3. Chein, I, An Introduction to Sampling, in Selltiz, C, et al, Research Methods in Social Relations, New York, Holt, Rinehart, and Winston, 1976.
4. Bartlett, EE, The Process of Patient Education in Medical Care, Physician's Patient Education Newsletter, 1(3): 1-2, October, 1978.

Health Education: Panacea, Pernicious or Pointless? - Revisited

Ellen J. Cohen, Ph.D.
New York City Board of Education
New York University Medical Center

As a consequence of the frustrations spawned by the escalating costs and the low returns of disease-oriented medicine, health education has begun to be touted as an antidote to the nation's health ills. The recent passage of the National Consumer Health Information and Health Promotion Act, coupled with educational programs sponsored by other HEW agencies, should increase governmental outlays for health education to $100 million annually. In the private sector, insurance companies have indicated a willingness to reimburse for health education.

Evidence that health education is effective and therefore worthy of these prodigious expenditures, is lacking. On the contrary, most of the empirical data point to the ineffectiveness of both general health education as well as patient education programs.

Despite recent extensive health educational campaigns to alter American smoking, driving, leisure, eating and drug habits, none have produced any notable successes. The millions of dollars spent on anti-smoking campaigns tend to refute the claim that inadequate funding is primarily responsible for these failures.

Since a wide discrepancy exists between health knowledge/attitudes and health actions, those who claim health education campaigns have produced the recent reductions in animal-fat consumption and in cigarette smoking among men are arguing by asseveration: saying things positively makes them so. There are no scientific data justifying the conclusion that education has been primarily responsible for these changes. Even the most ardent health-education advocates have acknowledged the effect of taxation and legislation on reducing cigarette consumption. Furthermore, the reduction in male cigarette smoking probably stems from the availability of profitable alternatives (e.g., pipes, cigars), whereas no such alternatives exist for women. Similarly, with animal-fat consumption, higher beef and dairy prices and the availability of alternatives may be the principal etiologic agents. Interestingly enough, teenage girls have been subjected to the most systematic and intensive educational programs, and they have shown an increase in cigarette consumption over recent years.

Haggerty observed that a careful review of the literature leaves one with only a few studies meeting two criteria: that alterations in behavior rather than knowledge or attitudes should be the end result of intervention; and that rigorous research methods with controls or at least good comparisons should be used. Haggerty argued that "when such rigorous criteria are used, few health education programs have shown much effect on behavior."

Our own analysis of all empirical studies reported in 11 major health-education journals during the three-year period 1975-77 indicated that two fifths of the studies had inadequate designs, one fifth yielded insignificant results, and one third showed statistically significant results but only minimal improvements (i.e., although the experimental group differed from the control, the resultant scores were below the 75 per cent levels of desired cognition, attitude or behavior). Only 7 per cent demonstrated statistically significant results with major improvements.

Other excellent reviews by Berberian et al., Green, Hosay and Young have similarly presented a rather bleak picture of health education's ability to alter behavior meaningfully.

Health education has shown some promise in the area of patient education. But most successes generally have been achieved in situations in which patients had procedures clarified or were taught self-administration technics -- e.g., diabetes, hemophilia or chronic renal disease. Patient education has been less successful in altering habits and life-styles. Health education combined with clinical services has also produced more sanguine results. However, no attempt has been made to separate the effects of additional treatment from educational influences.

Although the exponents of health education readily adknowledge the need for more carefully controlled studies, they nevertheless maintain that there is already "significant evidence" that educational programs can reduce health costs. Another common rejoinder has been to argue that there will never be a time when "all the data are in," since the context of society is constantly in flux. It therefore follows that funding must be committed in advance of any definitive conclusions regarding health education -- if such a state is ever attainable.

It is the contention of this paper that the inability of health education to alter behavior appreciably does not stem from the "primitive state of the art," but rather from contradictions inherent in the concept of health education as it exists in our society.

The basic philosophy of health education entails an individual solution to all health problems. Characteristic of this philosophy is the NIH/American College of Preventive Medicine task force statement that "...good health depends on individual knowledge, individual responsibility, and individual participation..."

The contradictions of the individualistic solution are illustrated by the following evidence: (1) Those with the greatest number of major life events such as frequent moves, shifting jobs, death of loved ones have the highest morbidity/mortality rates. (2) Stress resulting from these life events prevents Americans from practicing good health behavior. (3) The four major causes of death have been found to have a strong environmental component.

All of the above are not subject to individual control; therefore, changes in individual life styles will not remedy the problem.

In conclusion, although health education per se is not without merit, at this historical moment the possibilities for meaningfully influencing health through education and individual decision-making have been severly curtailed.

Getting Clear on What Health Education Is and Isn't

Wendy Diane Squyres, Ph.D.
Kaiser Permanente Medical Group
Oakland, California

Much of the controversy regarding the efficacy of health education is imbedded in misconceptions about what it is and what it should accomplish.

An overview of selected definitions of health education and patient education will be presented. A standard for defining health education, as well as measuring health education outcomes will be discussed. A review of patient education research and evaluation literature will be summarized, and case studies will be selected and measured against the standard. The utility of such an approach for the training of health professionals will be presented.

SYMPOSIUM

THE CLINICAL PERFORMANCE OF MEDICAL SCHOOL GRADUATES: THEORETICAL, METHODOLOGICAL, AND EMPIRICAL PERSPECTIVES

Organizer: Louise Arnold, Ph.D.

Chairman: Jerry A. Royer, M.D.

Participants: Sarah M. Dinham, Ph.D.
Christel A. Woodward, Ph.D.
Louise Arnold, Ph.D.
Jonathan Keck, Ph.D.
T. Lee Willoughby
Virginia Calkins
Jerry A. Royer, M.D.
John C. Reid, Ph.D.
Jon Veloski

Discussant: Rosalie E. A. Paiva, Ph.D.

THE CLINICAL PERFORMANCE OF MEDICAL SCHOOL GRADUATES: THEORETICAL, METHODOLOGICAL, AND EMPIRICAL PERSPECTIVES

Louise Arnold, Ph.D.
The Medical School, University of Missouri-Kansas City

Theoretical and methodological issues in follow-up investigations of the clinical performance of medical school graduates will be explored during this symposium. Its objectives are to explore the value of such studies and to indicate how they can be accomplished.

By providing a historical overview of follow-up studies, the first presentation will set the research context for studies of postgraduate clinical performance. It will note the purposes of these studies and describe the useful theoretical underpinnings available to guide such work. The second presentation, based on a survey of relevant research currently underway in North American medical schools, will discuss a number of methodological decisions which must be made in designing studies that utilize rating scales, the most common technique used to measure clinical performance of graduates. The third and fourth presentations, as in-depth reviews of work at three different medical schools, will illustrate what can be learned from studies of postgraduate clinical performance. Finally, a discussant will assess the symposium offerings in relation to the present "state of the art" of measuring postgraduate clinical performance and will indicate additional lines of inquiry.

The symposium topic is perennially timely, as medical schools and residency programs, in evaluating their educational processes, strive continually to examine the quality of patient care which their graduates provide.

HISTORICAL AND THEORETICAL PERSPECTIVES ON FOLLOW-UP INQUIRY IN MEDICAL SCHOOLS

Sarah M. Dinham, Ph.D.
The University of Arizona College of Medicine

The first focus in this introductory discussion is a retrospective view -- a historical perspective on follow-up study in medical schools. This historical review describes the array of follow-up studies in two ways: by examining the questions being posed in these studies, and by addressing the purposes of those studies. The second focus of the discussion is a prospective view -- an examination of evaluation models to see whether any suit our purposes and might therefore give structure to our inquiry.

Historical Review

The types of questions posed in follow-up studies are as varied as one could imagine. For example, a significant number of follow-up studies investigate relationships between student demographic characteristics/admissions decisions/curricular patterns and postgraduate variables (e.g.Graham and Otis, 1973). Further, follow-up research often concerns residency performance (e.g. Woodward, 1978; Dinham and Barbee, 1978). Then,some follow-up studies concern aspects of the graduates' practice of medicine -- from demographic concerns such as specialty choice and location (e.g. Yancik, 1975; Sachs, 1977; Stone and Knopke, 1977) to personal variables such as satisfaction or stress (e.g.Mawardi, 1979) to direct measures of clinical proficiency (Woodward, 1978). And as an aside, follow-up research often includes some statistical hide-and-seek to keep investigators entertained while we wait for late-returning questionaires (e.g. Hutchins, 1962; Anderson, 1974). Last but not least, a few studies cover the entire range; chief among these has been the massive AAMC Longitudinal Study of the Class of 1960.

Even this superficial review of the variety in follow-up studies raises more questions than it answers. Why are schools from Philadelphia to Tucson mounting impressive campaigns upon their graduates? Purposes of follow-up studies vary; a convenient descriptor of that variance is classification of studies (or of hypotheses or questions) as having descriptive, evaluative, or research purposes (Hain and Dinham, 1977).

Almost all follow-up studies have a descriptive or census component. My legislature wants to know how many of our graduates practice in Arizona, and their location, and what proportion are providing primary care; your Board of Trustees may want to know about your graduates' research productivity. Even more complex studies can be classified as descriptive -- for example Mawardi's (1965) monumental effort to trace and personally interview 200 twenty-year alumni of the innovative Western Reserve curriculum. The descriptive opportunities presented by the AAMC Longitudinal Study (Hutchins, 1965) provide another example. Not many follow-up studies can be called "research"; few, indeed, test causal hypotheses by comparison of randomized treatments. Hypotheses positing causal relationships have been examined, however, in many ex post facto (e.g. Schumaker, 1963; Yancik, 1977) and correlational (e.g. Tonesk, 1975) studies. In general, however, the possibilities for identifying causal relationships are few because of logistical constraints, and because the methodological pitfalls are many (Neufeld, 1978; Rothman, 1972).

By far the greatest number of follow-up studies are evaluation studies in which the purpose is to link findings directly to judgements of value and/or to decision making (e.g. Boulton and Johnson, 1970). Veloski's contribution to this symposium will illustrate this principle nicely.

Evaluation Models

Given that the most common purpose of follow-up study is evaluation, this presentation turns next, although briefly, to a question about evaluation itself. Does the burgeoning field of educational evaluation give direction to investigators pursuing follow-up inquiry? Which models of educational evaluation suit the needs of a medical school exploring aspects of its own outcomes?

Typologies and classifications of evaluation models are abundant (e.g. Worthen and Sanders, 1973; Hastings, 1978). From the traditions of formal accreditation to the nuances of "connoisseurship", evaluation models are formulated as explanations for relationships among needs, methodologies, findings, and audiences. Mercifully, medical educators conducting evaluative follow-up studies need choose from a more limited spectrum.

Several evaluation models receiving current attention seem very well suited as a theoretical base for follow-up inquiry. Decision making models (e.g. Stufflebeam, 1971) are useful in focusing attention on the actual uses for the evaluation. Behavioral objectives models are useful to the extent that a school has specifically defined its intended outcomes. The goal-free approaches advocated by Scriven (1973), far from being casual, can conceivably be beneficial to medical faculty not entirely sure what all the curriculum's outcomes might be. And the art/criticism models (e.g. Eisner, 1977) may best appeal to the sensibilities of those who believe medicine to be an indescribable art, even (perhaps especially) at the postgraduate level.

The theoretical perspective, then, is as rich and promising as the historical perspective is varied. Follow-up inquiry is alive and well, although challenging and complex, as this symposium's following presentations will show.

METHODOLOGICAL ISSUES IN STUDIES OF GRADUATE PERFORMANCE

Christel A. Woodward, Ph.D.
The Medical School, McMaster University

Increasingly, medical educators have looked beyond graduation to obtain information on the performance of their medical graduates as interns. This information is sought for the following reasons:(1) to examine the relationship among performance in medical school, in internship and on licensing examinations,(2) to plan curriculum change and/or monitor its impact,(3) to obtain feedback on admissions policies,(4) to evaluate the internship performance of graduates against pre-determined standards,(5) to compare the internship performance of graduates of one medical school with graduates of other medical schools.

Rating scales have been called the "backbone of evaluation in clinical education" (Levine & McGuire, 1971). They have several attractive features, and the problems characteristic of rating scales are less glaring when the data are to be treated in an aggregated fashion rather than be used to make decisions about individuals. These comments do not imply that rating scales should be the only way graduate performance is measured. Rather, because of difficulty in implementing other types of measures across numerous training sites, they are the most widely used measuring tools.

Purpose

This paper will focus on the methodological considerations in planning a study to obtain internship performance ratings. Both issues in study architecture and the design of rating forms will be highlighted by reviewing six studies completed or in progress at North American medical schools.

Issues in Study Architecture

When implementing a study, a number of methodological decisions must be made regarding its architecture. These decisions include: (1) what is to be rated, (2) when the rating will be made, (3) how the rater is chosen, (4) how many raters are chosen and (5) who completes the rating. The decisions made are somewhat different in each internship performance study reviewed. Little is known, however, about how each of these factors affects the performance data obtained.

Issues in Rating Form Design

Studies of internship performance reported in the literature and in progress differ not only in the dimensions outlined above. They have employed differing types of rating. The number of content areas assessed, type of content and the number of items used to assess a content area differ widely. Some forms include open-ended questions, while others allow only closed-ended responses. Interval scales, frequency scales, norm referenced scales and criterion referenced scales have been fielded. Most investigators report that experience with other interns is used as the standard by the rater, sometimes despite instructions to the contrary.

Other Methodological Issues

Clear instructions, simple format and content relevant to the internship setting are important considerations in obtaining useable data. Thus, pretesting the rating scale developed and/or using the first year of data collection for a pilot study are strongly recommended strategies. Current studies pay more attention to obtaining evidence of the reliability of the form used than earlier studies. Several groups have attempted to assess the content, construct, concurrent and/or predictive validity of the rating scale employed.

Most studies report response rates of 60-75%, although the sources of non-response may differ or not be indicated. Non-response potentially creates biases. If weaker graduates do not enter postgraduate training or refuse participation in the study, a positive bias may occur. Differences in compliance with the rating task at training sites affiliated with the graduates' medical school and at distant sites may bias the results obtained.

One of the major problems in analysis of data received from a large number of internship sites is that differing aspects of performance may not be equally valued at all sites and/or not be equally observable by the rater. The ratings given then may be more a reflection of the site than the intern's performance. This may also produce non-response and/or substitution of a different (home-grown) rating form.

Several additional problems that plague the analysis of data obtained using rating forms will be discussed.

Summary

It is clear that no absolute methodological standards exist in graduate performance studies either for the rating scale applied or in the architecture of the study. This paper has reviewed the diversity and the methodological questions which must be addressed in planning a study. Further work to find out *what* can be rated in and/or is seen as important to different training programs is recommended. Greater attention should be paid to experimenting with structural aspects of the study, examining potential sources of bias and their impact on the results obtained. This will allow increasing methodological sophistication and promote greater standardization of structural decisions.

CLINICAL PERFORMANCE OF GRADUATES AS VALIDATION OF SELECTION AND PROMOTION PROCEDURES IN TWO MEDICAL SCHOOLS

Louise Arnold, Ph.D., Jonathan Keck, Ph.D., T. Lee Willoughby,
Virginia Calkins, Jerry A. Royer, M.D., and John C. Reid, Ph.D.
The Medical School, University of Missouri-Kansas City
and
The Medical School, University of Missouri-Columbia

Criteria which medical school admissions and promotions committees use have typically been validated against measures of student performance during undergraduate medical education (Cuca et al., 1976). Relatively few investigators have validated these criteria against measures of performance of graduate physicians although such outcome measures might be viewed as the most essential for evaluating medical education (Wingard & Williamson, 1973; Barro, 1973).

This study utilizes the clinical performance of graduates in residency as the validator for selection and promotion criteria. It examines the questions of whether a combination of cognitive and noncognitive criteria used in selection and promotion accounts best for the variance in graduates' clinical performance and whether the same combination offers the best explanation for variance in performance of graduates from two medical schools, one with a four-year curriculum, the other with a six-year program.

The first phase of this work was conducted at the six-year school at the University of Missouri, Kansas City. Its 1978 graduates were evaluated during their first postgraduate year by supervisors of their residency programs. Their performance was rated with the Resident Evaluation Form (REF) which contains 33 bipolar scales that yield an overall assessment of clinical performance in nine categories (Keck and Arnold, in press). The total REF score was the dependent variable used here.

Independent variables, taken from the selection and promotion process at the School, were divided into two classes: cognitive and noncognitive. Cognitive variables consisted of undergraduate GPA; scores on an internal examination of medical facts and concepts; scores on the National Board of Medical Examiners' Examination, Parts I and II; instructors' evaluations of students' clinical performance in Year Four internal medicine, a Year Five elective related to the graduate's area of specialization, and Year Six internal medicine; and a peer evaluation of Year Six internal medicine. Noncognitive variables included a Year Six internal medicine self-evaluation, scores from Holland's Self-Directed Search, the graduate's gender, and ratings from a physician's admission interview. Previous research provided the justification for the above classification of variables as cognitive or noncognitive (Keck, et al., in press).

Stepwise multiple regression analyses were performed on each class of variables and on all the variables combined. Cognitive variables resulted in a corrected $\hat{R}^2$ of .27. Noncognitive personal characteristics, when entered into the multiple regression, produced an $\hat{R}^2$ of .22. Cognitive and noncognitive variables together produced an $\hat{R}^2$ of .51.

These findings agree with previous investigations of the clinical performance of UMKC graduates (Keck, _et al_., in press). Together, they indicate that a combination of cognitive and noncognitive variables used in both selection and promotion explains the variance of graduates' clinical performance much better than any individual variable or specific class of variables; and they support the practice of using personal attributes in selection and promotion of students.

In the second phase of the study, the results obtained at UMKC will be cross-validated at the Medical School, University of Missouri, Columbia, whose 1978 graduates were also rated with the REF by supervisors of their residency programs. A similar approach to data analysis will be followed. Results of the cross-validation will be presented, and comparisons of the validity of selection and promotion criteria for each type of medical education program will be drawn.

USING POSTGRADUATE CLINICAL PERFORMANCE TO MONITOR CHANGE IN THE MEDICAL SCHOOL

Jon Veloski
Office of Medical Education, Jefferson Medical College

Since 1969 the Jefferson Medical College of Thomas Jefferson University has been conducting a longitudinal study of its students and graduates. Data related to the students' personal background and interests at time of admission, evaluations of performance made during medical school training, attitudes and career plans at graduation, and measures of competence during the first year of postgraduate training have been collected. Using, for example, the data base on the clinical performance of twelve graduating classes during residency, it has been possible for the faculty to monitor the effects of changes in the medical school program. This presentation will include examples of methods and findings taken from longitudinal evaluations of three changes made in the medical school: an accelerated premedical program, the matriculation of increased numbers of women, and a change of the medical school curriculum. It is expected that these examples will demonstrate the types of questions which can be addressed when longitudinal data are readily available.

The evaluation of the accelerated program compared the clinical performance of a specified group of students (in the accelerated program) with those who followed the traditional program. The study of women graduates compared their postgraduate clinical performance with that of a sample of men matched for year of entrance, type of premedical program, and ethnicity. The study of the change of curriculum compared classes graduating prior to the year of curriculum change with those graduating afterward on measures of performance in postgraduate training. It was proven that the curriculum change produced some of the results which had been anticipated.

There is a need to monitor the results of medical education in order to assure that changes produce the expected results, and an ongoing longitudinal study can provide readily retrievable data which can be used to examine those results. This presentation will describe how one medical school has been able to do this.

SYMPOSIUM

THE ROLE OF DECISION ANALYSIS IN MEDICAL EDUCATION AND PRACTICE

Chairman:	G. Anthony Gorry, Ph.D.
Participants:	Jerome P. Kassirer, M.D. Harvey V. Fineberg, M.D. Stephen G. Pauker, M.D. Arthur S. Elstein, Ph.D. John C. Peirce, M.D.

THE ROLE OF DECISION ANALYSIS IN MEDICAL EDUCATION AND PRACTICE

G. Anthony Gorry, Ph.D.

Program for Health Management
Baylor College of Medicine

Overview:

The objective of this symposium will be to delineate the role that certain formal techniques of decision-making, collectively called decision analysis, can play in the education of young physicians. The participants in the symposium have had experience in teaching decision analysis to medical students, house officers and practicing physicians, and each has made contributions to the analysis of clinical decision making. In their talks, the participants will describe the techniques of decision analysis, and they will endeavor to assess realistically the potential value of the use of decision analysis in medical education and practice.

In view of the essential role that clinical judgment plays in the practice of medicine, it is remarkable that in medical education so little attention is paid to the nature and acquisition of these skills. During the past fifty years, medical science has constructed an extensive factual base upon which to anchor the practice of clinical medicine. With each day, new means are developed to probe more deeply into the body - to see more clearly at an early stage in its development, the pathologic anatomy that underlies the clinical presentation of disease. But despite the advances of medical science, clinical practice remains a far more complex undertaking than the routine application of diagnostic and therapeutic technology to the patient's problem. The clinical setting is one in which ambiguity, not certainty, is the rule, and one of the most highly valued clinical skills is the ability to exercise those intellectual qualities that collectively are termed clinical judgment. In contrast to the expanded scientific language of medicine, the language used for describing the mental processes of clinicians is grossly underdeveloped. Although much of what is termed clinical judgment will always be the product of experience, there may be clinical problem-solving strategies which can be made explicit, if physicians are equipped with the proper language for describing them - a better language for thinking about thinking.

Increasingly the medical literature contains articles which describe formal aids to clinical decision making. One of the most important of these aids is decision analysis - a framework for making decisions in the face of uncertainty. Decision analysis does not seek an increased understanding of the pathophysiological mechanisms of disease as an end in itself. It is concerned with the diagnostic and therapeutic choices available to both physician and patient, and in ways in which these choices can be made more consistently with the outcomes desired. Decision analysis focuses attention on questions of uncertainty, chance, ambiguity, and values inherent in almost every clinical situation of any consequence.

Sufficient experience has been gained in recent years to suggest that in certain instances knowledge of the techniques of decision analysis can be of considerable benefit in clinical practice and in the education of young physicians. Because of the more precise nature of these methods, they provide a more explicit language for describing and discussing certain kinds of intellectual problems in the clinical setting. These problems include interpreting of evidence, determining the value of information and balancing risks against expected benefits. Additionally, with the techniques of decision analysis it is possible to assess the attitude of the patient towards the possible consequences of medical actions which may be taken on his behalf. In view of increasing social concern about the quality and efficiency of medical decision making, the consideration of more formal approaches to these problems in the medical curriculum seems particularly timely.

THE USE OF DECISION ANALYSIS AS A TOOL TO TEACH "CLINICAL JUDGMENT"

Jerome P. Kassirer, M.D.

Tufts University School of Medicine
New England Medical Center Hospital

Clinical judgment is a complex process in which an enormous body of facts is utilized along with certain decision-making processes to arrive at good decisions regarding patient care. Medical education has emphasized the acquisition of facts but has paid little attention to the principles of decision-making. In part, this deficiency in the education process stems from a lack of explicit methods for clinical decision-making. Most clinical teachers teach by example, and their attempts to provide explicit guidance generally yield rules of procedure. For example: one should carry out a pulmonary arteriogram on a patient suspected of having a pulmonary embolus. Unfortunately, as any clinician knows, patients do not always fit into neat and simple, typical cases. In the great majority of instances, many complicating factors influence the clinical picture and, prevent one from using a set rule or procedure. Decision analysis can help in this regard, because it embodies an explicit approach to making complex clinical decisions and makes it possible for us to teach the elements of clinical decision-making to students, to house officers, and to physicians.

Many of the elements of clinical decision-making can be illustrated by considering the process a skilled physician might utilize in thinking about how to deal with a specific patient. Suppose a patient is encountered who had a splenectomy a month earlier and now shows clinical findings suggestive, but not diagnostic of a subphrenic abscess. The physician might consider many factors before deciding whether or not to operate. How likely is it that an abscess is present? Could the clinical findings be explained by a medically treatable disorder, for example, pancreatitis? If the patient is subjected to surgery, will a surgical or anesthetic complication occur, and how serious will such an event be? If surgery is carried out, will the surgeon be able to deal successfully with the abscess? How serious will it be if the lesion is not adequately excised or drained? What are the consequences of failing to operate if the patient has a surgically correctable lesion? Would a surgical lesion resolve under medical therapy? If so, how often would this occur and how often would complications occur? How serious would these complications be ?

The integration of these diverse considerations into a reasoned decision is not a simple matter. The integration process is usually carried out in tacit fashion, largely because most physicians do not have command of an explicit method for therapeutic decision-making. Such an explicit decision-making method will be described here.

Decision analysis requires the explicit identification of the choices open to the clinician. These choices may be to treat or not to treat, or between one treatment and another or many others, or between a test and a treatment. Decision analysis requires that the potential outcomes of each choice be designated specifically. For a test, the accuracy of the test and the risk associated with it must be assessed explicitly; for a treatment, the outcomes that may result must be defined and probabilities must be attached to these outcomes, and an assessment of the value, or utility, of each outcome must be made. Certainly the outcome for a patient who suffers a serious complication of treatment and derives no benefit from treatment has the lowest utility, whereas the outcome for a patient who receives appropriate treatment and is cured of the disorder has the highest utility.

These elements of the decision-making process, namely the specification of available choices, the explicit description of the outcomes of each choice, and the assignment of a probability and utility of each outcome defines the explicit approach to decision-making known as decision analysis. Once numerical values are assigned to the probabilities and utilities of each outcome, it is straightforward to calculate an "expected value", a numerical assessment of each choice available to the clinician. The choice with the highest expected value is the one with the greatest likelihood of a successful outcome.

Our experience in using decision analysis in the teaching of medical students and housestaff indicates that it has considerable potential as a framework within which the teacher can organize his discussion of clinical decision-making and within which the student can better appreciate the factors weighed by clinical judgment. Using decision analysis as a guide, the teacher can analyze the decision made in a particular case, but more important, he can analyze the decision-making process itself rather than only the facts that make up the elements of the decision. Choices can be examined, outcomes can be reviewed, and probabilities and utilities can be reassessed to sharpen the students' awareness of the issues involved and how those issues bear on the case in question. In all, for many clinical problems, decision analysis supplants an implicit, vague decision-making process with a mechanism that is clear and explicit and seems readily applicable in the teaching of clinical judgment.

THE EXPECTED VALUE OF DIAGNOSTIC TESTS

Harvey V. Fineberg, M.D.

Center for the Analysis of Health Practices
Harvard School of Public Health

Decision Analysis provides a useful framework for assessing the expected value of diagnostic tests. Tests discriminate, usually imperfectly, among different diagnostic possibilities. From a decision-making point of view, the near term purpose of a test is to revise probability estimates of the presence or extent of disease. The longer term purpose is to improve the choice of therapy and health outcome. Along the way, a test may prompt or discourage performance of other tests. Performance of a test consumes resources itself and may also induce or save consumption of additional resources.

The expected clinical value of test information is the difference between the expected health outcome with the test and that without the test. Measuring the expected value of a test requires specifying and valuing the health outcomes of interest and identifying the initial and revised probabilities of disease conditional upon various test results. A decision tree is a convenient way to display these features. In order for a test to be of net positive value, it must satisfy a necessary and a sufficient condition. The necessary condition is that at least one decision following the test must depend on the results of the test; in other words, at least one possible test result must lead to the selection of a subsequent choice which is different from what would have been selected in the absence of the test. The sufficient condition is that the average benefit from this change in decision must outweigh any intrinsic risk of using the test.

The expected value of a test thus depends on characteristics of the test, the characteristics of the persons tested, and characteristics of the possible diseases and their treatments. Both positive and negative test results may have net positive value; what counts is the extent of change from previous estimates of disease probability. The expected value of a test that discriminates between two possible diseases, each with a preferred treatment, is greatest when the prior probability of disease is at the point where (without using the test) each treatment would have offered the same expected outcome.

The expected value of multiple tests is much more complicated than for single tests, in part because probabilities of disease given various sequences and combinations of tests must each be estimated. Various test sequences may entail different risks and incur different resource costs. The considerations in choosing an optimal sequence of tests include: 1) the time frame for decision making, 2) the time required for each test, 3) interaction (interference/enhancement) of

446

one test with another, 4) risk to the patient, 5) resource costs and 6) information obtained.

The diagnostic information contributed by multiple tests is greatest when test results are conditionally independent. When tests are not conditionally independent, they provide aggregate diagnostic information which is less than might be anticipated, and probability revision requires assessment of the conditional probabilities of each combination of test results given each disease state.

Acting as if tests were conditionally independent, when they are not, and liberalism and conservatism in revising probabilities using multiple tests, are among the factors leading to excessive testing. Medical students and physicians can learn to apply principles of decision analysis to the problem of choosing medical tests, and we believe that more widespread teaching of these ideas would contribute to more rational medical decision making.

These points will be illustrated by reference to tests in several clinical situations, including the evaluation of patients with chronic liver failure and of patients with hypertension.

THE USE OF DECISION ANALYSIS TO INCORPORATE PATIENT ATTITUDES INTO CLINICAL DECISION MAKING

Stephen G. Pauker, M.D.

Tufts University School of Medicine
New England Medical Center Hospital

A medical decision should reflect the attitudes and desires of the individual patient with respect to the potential outcomes of a given diagnostic or therapeutic choice. In general, however, the attitudes of patients are only implicitly acknowledged, and physicians make those decisions which they believe to be best in the circumstances as they perceive them. Unlike the family physicians who in the past may have been very close to their patients, many consultants now make medical decisions for patients whom they hardly know. Our society is sufficiently heterogeneous that it is unwise to assume that the physician's attitudes necessarily are those of the patient's. Indeed, ignoring the patient's attitudes runs counter to the principles of informed consent and patient participation which now are emphasized.

Decision analysis offers a method for explicitly incorporating patient attitudes in the decision-making process. In analyzing a problem, the probability of each potential outcome is separated from the value or utility of that outcome. In general, the physician is better able to assess, the pertinent probabilities, but the patient can make better judgments of the utilities. The physician can use the formalism of decision analysis to calculate the expected value of each action and hence to identify the action which is best in light of the patient's attitudes.

The patient can be guided in providing these measures of his attitudes by use of a technique known as utility assessment. The patient then chooses the best and the worst outcome for him. For each intermediate outcome the patient is then asked to choose between the intermediate outcome for certain and a gamble between the best and the worst outcomes. The odds on the gamble are altered until a point of indifference is reached between the patient's assessment of the value of that gamble and the intermediate outcome for certain. The utility of the intermediate outcome can then be taken to be the percent chance of the best outcome at the indifference point. By repeatedly applying this technique one can rapidly assess the patient's utility for each outcome and this utility can then be used in the decision-making process.

We have had substantial experience in using utility assessment in two clinical situations; 1) the decision about coronary bypass surgery in patients with disabling angina pectoris and 2) the decision about amniocentesis facing prospective parents who are at risk for having a child affected by a congenital abnormality such as Triscomy 21 (Down's Syndrome). Patients accept the procedure readily and although the questions which they must answer are difficult, most patients appreciate that this approach allows them to participate more fully in the decision-making process.

Since we have found this approach useful in dealing with our own patients, we have recently begun offering a course to medical students which covers many aspects of clinical decision-making, including this method of incorporation of patient attitudes. The students have responded favorably to this course. As a part of the course, they were asked to assess each other's utilities as if they were patients facing the two decisions outlined above. There was sufficient heterogeneity in the class to demonstrate that one cannot make assumptions about the patient's attitudes, but rather one must explicitly assess his utilities. Certainly, there are many clinical decisions where patient attitudes may be sufficiently homogeneous and can be sufficiently well approximated by the physician's own values so that such explicit debriefing of patient attitudes may not be necessary. There remains, however, a rather broad class of clinical problems where consideration for the patient's own values is not only appropriate but may be mandatory. In those situations the use of formal decision analysis to separate the complex decision into 1) assessment of the values of the potential outcomes to the patient and 2) the estimation by the physician of the likelihood of each potential outcome in the particular patient can help them jointly make good choices. We have shown that these techniques can be taught to students. We believe that those students who have been exposed to this approach should have a basis for more explicit and rational decision-making and indeed may show greater sensitivity to variations in patient attitudes toward medical care.

EXPLORING VALUES AND VALUE CONFLICTS VIA DECISION ANALYSIS

John C. Peirce, M.D. and Arthur S. Elstein, Ph.D.

Department of Medicine and
Office of Medical Education Research and Development
Michigan State University, and
St. Mary's Hospital, Grand Rapids, Michigan

Clinical decision making is often difficult because the problems are so complex, with so many aspects to be considered, and so much information potentially available. Information overload can impede good decision making, particularly for novices. It is important for medical students and inexperienced physicians to have a clear picture of the overall structure of a problem and to understand the central issues at stake in a decision. Structuring a decision tree is a particularly useful way of higlighting the key issues. Decision trees and decision analysis incorporate two features of most clinical decision making: They can represent a time-sequence of decisions and they must usually rely upon subjective estimates of certain parameters. In decision analysis, these parameters are probabilities and utilities. The purpose of this presentation is to show how the decision analytic perspective on clinical medicine employs the behavioral and social sciences in ways that are not yet widely recognized and appreciated.

Thus, there are three major components to an analysis; structuring the tree to represent the problem, estimating path probabilities, and assessing the utility of each outcome. Our central concern here is the value or utility of the outcomes specified in the decision tree. By distinguishing between knowledge about the likelihood of particular outcomes and ordered preferences for the outcomes, decision analysis encourages patients to become participants in making decisions that are best for them. The physician can serve as the patient's agent, not as the sole decision maker. This type of decision making is particularly relevant in chronic diseases which inevitably have a prolonged and sustained impact upon the patient. In assessing a patient's utilities, the physician deals necessarily with motivations, hopes, frustrations and fears. The physician must be able to define outcomes in terms that the patient can understand. Thus, empathic understanding and communication skills are emphasized within the context of an approach to rational decision making.

We are presently engaged in exploring the usefulness of decision analysis in helping patients decide about transplantation or hemodialysis as the preferred treatment for end stage renal disease. If a patient initially elects hemodialysis, we are concerned with a continual process of deciding whether to continue hemodialysis or to elect the irreversible alternative of transplantation. Because the patient's utilities may change over time, the problem for the physician and for the patient becomes not only translating values to a numerical scale, but sampling these values over time so as to update periodically the choosing of treatment. Other examples of this situation are deciding between medical and surgical management of peptic ulcer disease, chronic ulcerative colitis, and various neoplasms. In the medical management of these conditions, utilities may change over time and should be repeatedly assessed. We believe utility assessment can be used to ask the patient, "Where do you stand now?" Research is needed to determine how useful and meaningful this mode of affective exploration is to both physicians and patients and how sensitive it is to perceived changes in the desirability of alternative treatments.

The analytic distinction between probability and utility is also useful in making explicit some value commitments of which decision makers may be unaware. For example, using decision analysis, one of us (ASE) has retrospectively studied a controversial decision regarding treatment of a newborn with hemolytic jaundice. Two physicians agreed upon the diagnosis and etiology, but their treatment choices differed. One preferred immediate transfusion while the second preferred to continue phototherapy. Decision analysis disclosed that these two choices were predicated upon very different estimates of the probable mortality rate of transfusion and different utility assessments of the undesirability of possible severe and moderate CNS damage, two possible outcomes of delaying transfusion. The two physicians did not realize that these different estimates lay under the surface of their disagreement. Thus, decision analysis has considerable potential for elucidating the reasoning behind controversial decisions and for helping both physicians and patients become aware of their value commitments.

Decision analysis is a promising approach to several psychological questions relevant to clinical judgment. Consequently, it deserves the serious attention of medical educators preparing physicians for practice in the next decade. While decision analysis is sufficiently time-consuming that it cannot be routinely applied in all clinical situations, it has heuristic value in helping the learner focus on key issues. Thus, it can be employed as an educational device in circumstances where a more leisurely exploration of the complexities of a case is both desirable and necessary.

INDEX -- AUTHORS AND SYMPOSIA PARTICIPANTS

ALBANESE, Mark . . . 59
ALBO, Dominic . . . 107
ALEXANDER, F. . . . 67
APPLEGATE, William B. . . . 399
ARNOLD, Louise . . . 417
ARSHAM, Gary M. . . . 407
BAKER, George L. . . . 87
BAKEWELL, William E. . . . 145
BARTLETT, Edward E. . . . 407
BASHOOK, P. G. . . . 258
BAUMBER, J. S. . . . 67
BEALE, Pat . . . 252
BERNER, Eta . . . 377
BLAUSTEIN, Ernest . . . 317
BOSSHART, Donald A. . . . 45
BRITTON, James , III . . . 279
BROOKS, C. Michael . . . 252
BROWN, Sam . . . 299
BRYANT, G. D. . . . 205
BRYANT, H. . . . 67
BUDRICH, Margaret . . . 252
BUNCE, Jane V. . . . 185
BURG, Fredric D. . . . 105
CALHOUN, Judith G. . . . 197
CALKINS, Virginia . . . 417
CAMPBELL, E. J. M. . . . 214
CANDEE, Daniel . . . 93
CANNON, Nass J. . . . 299
CARTWRIGHT, Lillian Kaufman . . . 367
CASBERGUE, John . . . 219
CASSADY, George . . . 252
CHAFF, Sandra L. . . . 367
CHEATHAM, Claudia M. . . . 9, 145
CLAWSON, C. C. . . . 221
CLINE, David W. . . . 247
CLUFF, Leighton E. . . . 329
COHEN, Ellen J. . . . 407
COLOMBO, Sandra . . . 177
COLVIN, Ed. . . . 252
COOK, Charles D. . . . 93
CORMAN, Lourdes C. . . . 113
COWAN, Robert B., Jr. . . . 20
CUNNINGHAM, Russell D. . . . 299
CURRY, Lynn . . . 270
CURTIS, George M. . . . 20
DANIELS, M. Harry . . . 87
DAVIS, Wayne K. . . . 191, 197, 264
DEIGHTON, Bruce C. . . . 133
DEMAREST, Lawrence . . . 247
DICKOFF, Steven . . . 208
DIELMAN, T. E. . . . 191
DIENST, Evelyn R. . . . 329
DINHAM, Sarah M. . . . 337, 417
DONNELLY, Julie C. . . . 99
DONNELLY, Michael B. . . . 165
DOWNING, Steven M. . . . 124
DUVAL, Merlin K. . . . 407
EISENBERG, John M. . . . 399
ELLIOTT, Paul R. . . . 131
ELSTEIN, Arthur S. . . . 139, 345, 377, 427
ENGEL, John D. . . . 377
ENGELHARDT, H. Tristram, Jr . . . 85
ERVITI, Vivian . . . 185
ESSEX, Diane . . . 14, 305
EVANS, George P. . . . 159
FABREY, Lawrence J. . . . 185
FEIGHTNER, J. W. . . . 214
FICKEN, Roland P. . . . 291
FINEBERG, Harvey V. . . . 427
FLEISHER, Daniel S. . . . 165
FLETCHER, Richard . . . 355
FOSTER, Patricia Joan . . . 233
FREYMANN, John G. . . . 399
FRIEDMAN, Charles P. . . . 9, 145, 183
FUNDYTUS, D. . . . 67
GALLAGHER, Richard E. . . . 133
GARG, Mohan L. . . . 399
GARLAND, Michael J. . . . 399
GARRARD, Judith . . . 221
GARRISON, Jeffrey H. . . . 165
GELLER, Lester M. . . . 337
GERBER, Walter . . . 59
GJERDE, Craig . . . 177
GORRY, G. Anthony . . . 427
GROVER, Paul L., Jr . . . 389
HARASYM, P. H. . . . 67
HARPER, L. . . . 78
HARRIS, Ilene . . . 221, 227
HEINS, Marilyn . . . 367
HICKEY, James R. . . . 208
HINZ, Carl F., Jr. . . . 65
HISS, Roland G. . . . 264
HOCK, Robert A. . . . 32
HOLMAN, Gerald . . . 355
HUDSON, James I. . . . 399
HULL, Alan L. . . . 191, 197

HUSTED, Susan R. 93
JACKMAN, Mark F. 317
JACOBY, L. L. 214
JOHNSON, Barbara F. 47
JOHNSON, C. Warner 277
JORDAN, Ronald 355
KAPLAN, Stanley 165
KASSIRER, Jerome P. 427
KASTNER, Brian A. 208
KECK, Jonathan 417
KNOX, Eric 252
KROL, Ronald A. 20
LADUCA, Anthony 377
LANGDON, Lynn O. 345
LANIER, Roger A. 47, 171
MAATSCH, Jack L. 124
MACDONALD, Roderick 264
MACK, Lois P. 208
MACPHAIL, I. 67
MAISIAK, Richard 279
MALY, Betty Joan 285
MANDELBAUM, Dorothy Rosenthal 367
MANNING, Robert T. 355
MARKUS, John F. 239
MASHBURN, N. Kay 3
MAST, Terrill 159, 239
MATTHES, Sandra 59
MAXWELL, J. 258
MCDOLE, Janet 26
MCGAGHIE, William C. 377
MESKAUSKAS, John A. 345
MESNARD, Jane E. 208
MEREDITH, Ronald 279
MIDDLETON, Jayne 297
MOLIDOR, John B. 139
NEUFELD, Victor 73, 345
NIEBUHR, Bruce R. 151, 171
NORCINI, John J. 345
NORMAN, G. R. 73, 205, 214
PAGE, Brent 107
PAINO, Sharon E. 247
PAINVIN, C. 73
PAIVA, Rosalia E. A. 118, 417
PALMISANO, Paul A. 299
PATRIDGE, Mark I. 227
PAUKER, Stephen G. 427
PERKOWSKI, Linda 47
PETZEL, Robert A. 227
PIERCE, John C. 427
PORTER, Carol Q. 9, 145
POZEN, Janet T. 329
PRESHAW, R. 67
PRICE, Philip 107
PUHL, Carolyn 165
PUTNAM, R. Wayne 270
RATNER, Charlene 208
RAVIN, Arnold W. 317
RAVITCH, Michael M. 389
REID, John C. 417
RICHARDS, Ronald W. 389
RIPPEY, Robert M. 38, 78
ROBBINS, John 113
ROCK, Miriam B. 317
ROENNEBURG, Marcella 285
ROSE, Steven D. 113
ROSINSKI, Edwin F. 329
ROTHMAN, Arthur I. 157
ROYER, Jerry A. 417
RUGGILL, Jane 53, 337
RUTALA, Paul J. 53, 337
SABERS, Darrell 53, 337
SACKOFF, Judith 389
SANDLOW, L. J. 258
SANDSON, John I. 317
SCHERMERHORN, Gerry R. 389
SCOTT, C. W. 299
SCOTT, Lowell K. 299
SHATZER, John H. 26
SHEEHAN, T. Joseph 93
SHELLENBERGER, Sylvia 3
SHELTON, Steven R. 151
SILBER, David L. 159
SLOTNICK, Henry B. 389
SMITH, Douglas U. 133
SOLER, Norman G. 239
SORLIE, William E. 14, 305
SOSNOWSKI, M. 67
SPIRO, Joanna 285
SPRAFKA, Sarah A. 377
SQUYRES, Wendy Diane 407
STEWART, Peter A. 317
STILLMAN, Paula L. 53, 337
STIMMEL, Barry 337
STONE, Howard L. 203
STROSS, Jeoffrey K. 264
SUMNERS, Janes 252
TAYLOR, Calvin W. 107
TAYLOR, Cindy 355
THORUP, Oscar A. 245
TREADWELL, T. Walter 3
TRENT, Pamela J. 32
TURNER, Malcolm 252
VELOSKI, Jon 417
VOYTOVICH, Anthony E. 38, 78
VU, Christine Nu Viet 311
WADE, Kenneth J. 291

WALKER, I. 73
WATANABE, M. 67
WEBSTER, George D. 345
WEINBERG, Ethel 337
WERNER, Shirley A. 151
WHELAN, G. 73
WILE, Marcia A. 208
WILLIAMS, Byron 171
WILLIAMS, Reed G.. . . . . . .26. 159
WILLOUGHBY, T. Lee 417
WINGENFELD, Juanita R. 208
WIRTSCHAFTER, David D. 252
WOODWARD, Christel A. 417
WYSE, G.. 67
YANCY, W. Sam 32